US006529499B1

(12) **United States Patent**
Doshi et al.

(10) **Patent No.: US 6,529,499 B1**
(45) **Date of Patent: Mar. 4, 2003**

(54) **METHOD FOR PROVIDING QUALITY OF SERVICE FOR DELAY SENSITIVE TRAFFIC OVER IP NETWORKS**

(75) Inventors: **Bharat Tarachand Doshi**, Holmdel, NJ (US); **Enrique Hernandez-Valencia**, Highlands, NJ (US); **Kotikalapudi Sriram**, Marlboro, NJ (US); **Yung-Terng Wang**, Marlboro, MA (US); **On-Ching Yue**, Middletown, NJ (US)

(73) Assignee: **Lucent Technologies Inc.**, Murray Hill, NJ (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/158,694**

(22) Filed: **Sep. 22, 1998**

(51) **Int. Cl.**[7] **H04L 12/66**; H04L 12/26

(52) **U.S. Cl.** **370/352**; 370/230; 370/468

(58) **Field of Search** 370/351–356, 370/400–402, 229–238.1, 522–524, 395.2, 395.21, 395.41, 395.52, 468; 709/223–229; 320/209

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,914,650 A		4/1990	Sriram	370/60
5,434,852 A	*	7/1995	La Porta et al.	370/524
5,463,620 A		10/1995	Sriram	370/60
5,732,078 A	*	3/1998	Arango	370/355
5,751,712 A	*	5/1998	Farwell et al.	370/431
6,064,653 A	*	5/2000	Farris	370/352
6,078,582 A	*	6/2000	Curry et al.	370/352
6,094,431 A	*	7/2000	Yamato et al.	370/395.21
6,097,722 A	*	8/2000	Graham et al.	370/395.21
6,205,211 B1	*	3/2001	Thomas et al.	379/125
6,292,478 B1	*	9/2001	Farris	370/352

FOREIGN PATENT DOCUMENTS

GB	2317 308 A	3/1998	H04L/12/46

OTHER PUBLICATIONS

"A New ATM Adaptation Layer for Small Packet Encapsulation and Multiplexing"; John H. Baldwin, Behram H. Bharucha, Bharat T. Doshi, Subrahmanyam Dravida and Sanjiv Nanda; Bell Labs Technical Journal, vol. 2, No. 2, Spring 1997.

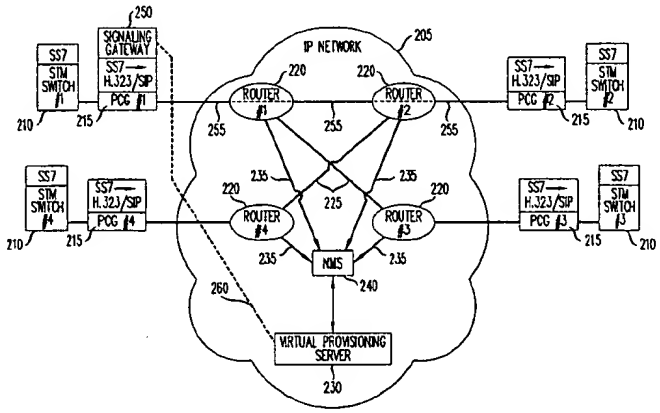
(List continued on next page.)

Primary Examiner—Steven Nguyen
(74) *Attorney, Agent, or Firm*—Troutman, Sanders, Mays & Valentine

(57) **ABSTRACT**

A quality of service guarantee for voice and other delay sensitive transmissions within an Internet Protocol (IP) network is provided by identifying the IP network path utilized for IP packet transmission between source and destination edge devices and virtually provisioning IP network path bandwidth for priority voice traffic. Priority for voice packets and admission control of new voice calls (and other delay sensitive traffic) based on the remaining available capacity over the IP network path guarantees that high priority voice (and other delay sensitive traffic) meet stringent delay requirements. A Virtual Provisioning Server is utilized to maintain bandwidth capacity data for each path segment within the IP network and to forward the bandwidth capacity data to a Signaling Gateway. The Signaling Gateway determines whether to accept or reject an additional delay sensitive traffic component based upon available bandwidth capacity for an IP network path. The Signaling Gateway then signals the originating source edge device as to its determination to accept or reject. Quality of Service guarantees concerning acceptable delay and jitter characteristics for real-time transmission over an IP network are therefore provided without the need to directly signal the individual IP routers over which an IP network path is established.

20 Claims, 5 Drawing Sheets



OTHER PUBLICATIONS

"Voice of ATM Using AAL 2 and Bit Dropping: Performance and Call Admission Control", Kotikalapudi Sriram and Yung-Terng Wang; Proceedings of the IEEE ATM Workshop, pp. 215–224, 1998.

"Anomalies Due to Delay and Loss in AAL2 Packet Voice Systems: Performance Models and Methods of Mitigation"; Kotikalapudi Sriram, Terry G. Lyons and Yung-Terng Wang; INFORMS Telecommun. Conf., Boca Raton, Fl. Mar. 8–11, 1998.

Kostas, T.J., et al., "Real-Time Voice Over Packet-Switched Networks", IEEE Network: The Magazine of Computer Communications, US, IEEE Inc., New York, vol. 12, No. 1, Jan. 1, 1998, pp. 18–27.

White, P.P. "RSVP and Integrated Services in the Internet: A Tutorial" IEEE Communications Magazine, US, IEEE Service Center, Piscataway, NJ, vol. 35, No. 5, May 1, 1997, pp. 100–106.

* cited by examiner

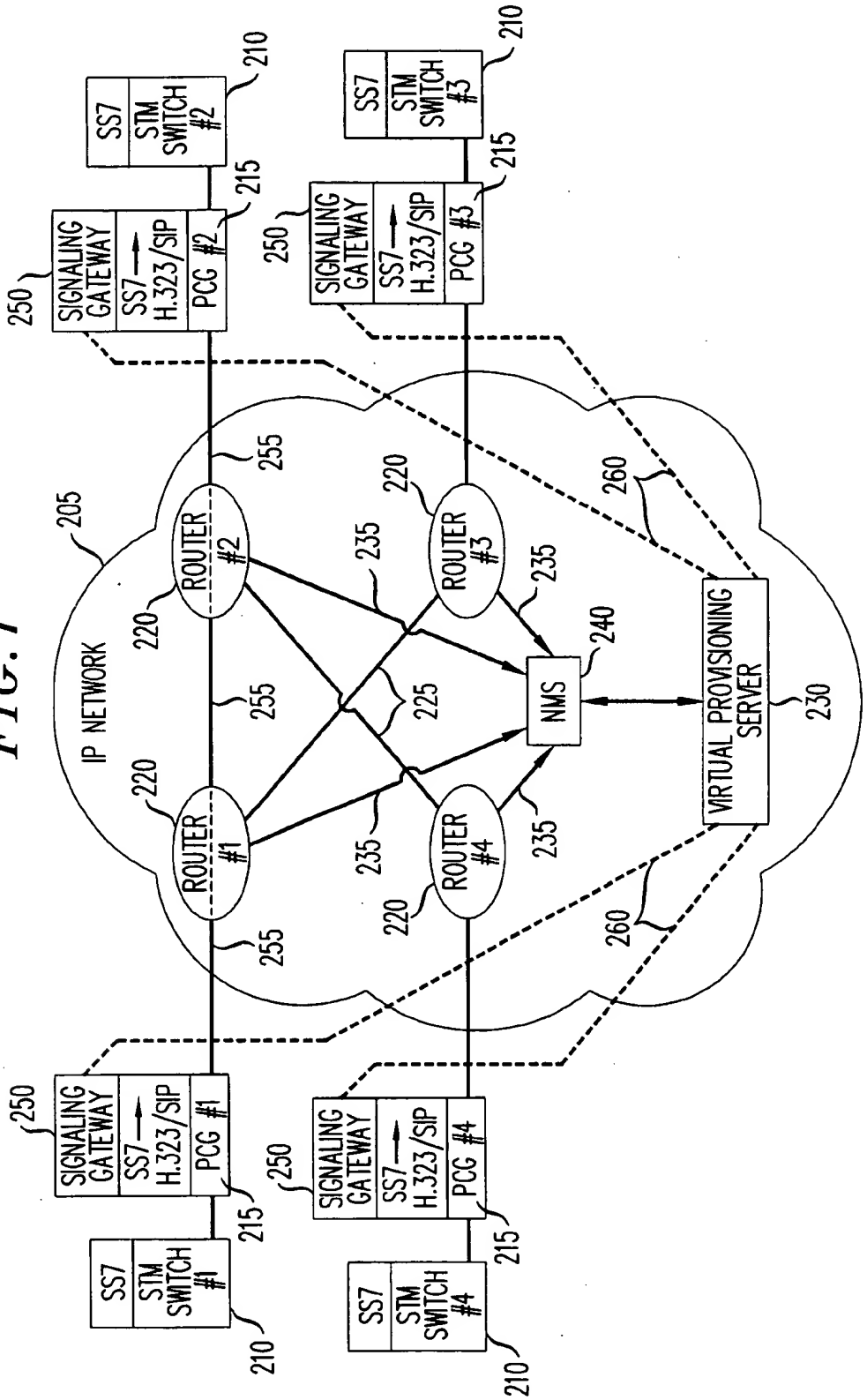

SS7 STM SWITCH #1
SS7 STM SWITCH #2
SS7 STM SWITCH #3
SS7 STM SWITCH #4
SIGNALING GATEWAY SS7→H.323/SIP PCG #1
SIGNALING GATEWAY SS7→H.323/SIP PCG #2
SIGNALING GATEWAY SS7→H.323/SIP PCG #3
SIGNALING GATEWAY SS7→H.323/SIP PCG #4
IP NETWORK
ROUTER #1
ROUTER #2
ROUTER #3
ROUTER #4
NMS
VIRTUAL PROVISIONING SERVER
205
210
215
220
225
230
235
240
250
255
260

FIG. 1

FIG. 2

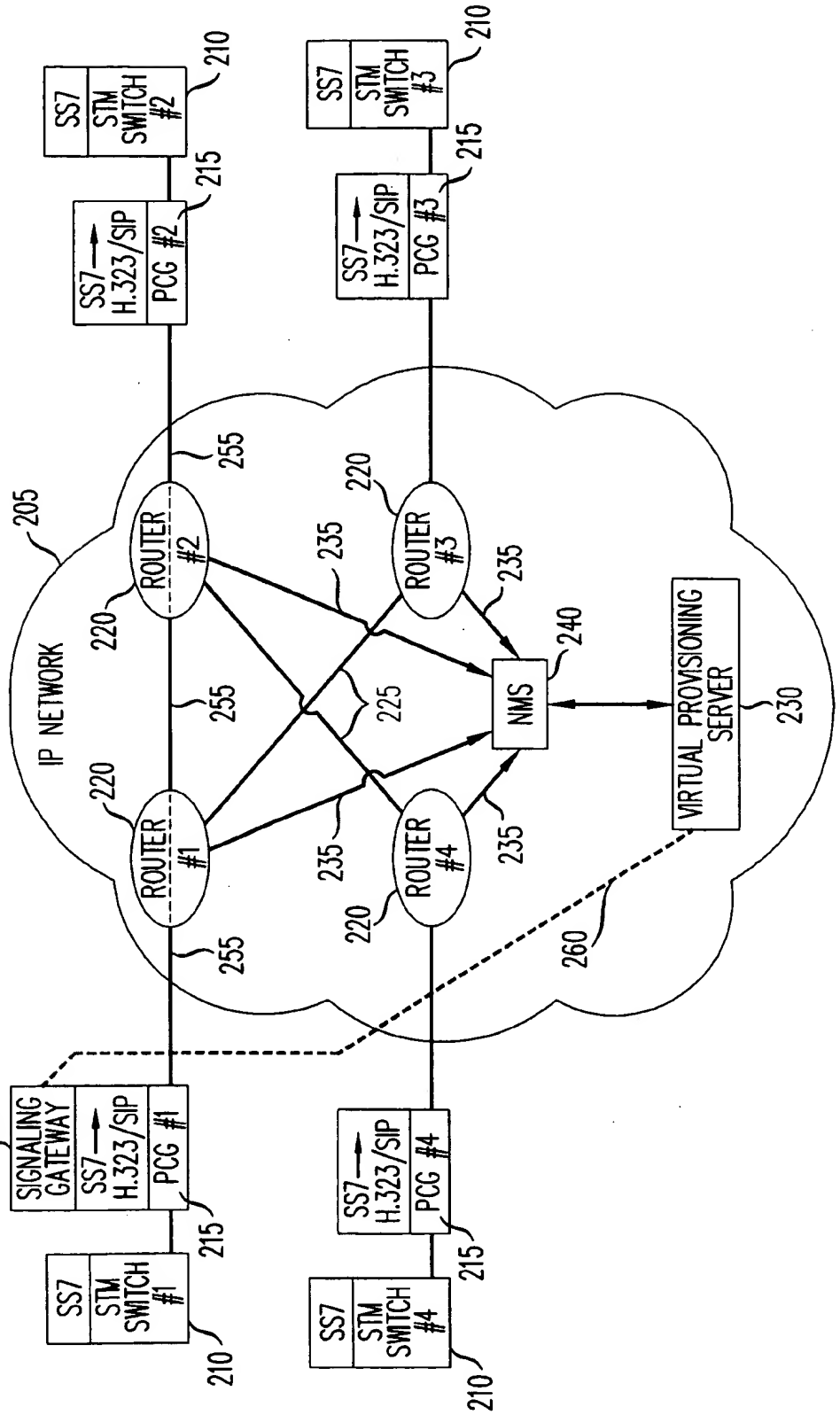

SS7 STM SWITCH #2
210
SS7 H.323/SIP PCG #2
215
SS7 STM SWITCH #3
210
SS7 H.323/SIP PCG #3
215
205
IP NETWORK
ROUTER #2
220
255
ROUTER #3
220
235
235
ROUTER #1
220
255
ROUTER #4
220
235
235
225
NMS
240
VIRTUAL PROVISIONING SERVER
230
260
255
SIGNALING GATEWAY
250
SS7 H.323/SIP PCG #1
215
SS7 H.323/SIP PCG #4
215
SS7 STM SWITCH #1
210
SS7 STM SWITCH #4
210

FIG. 3

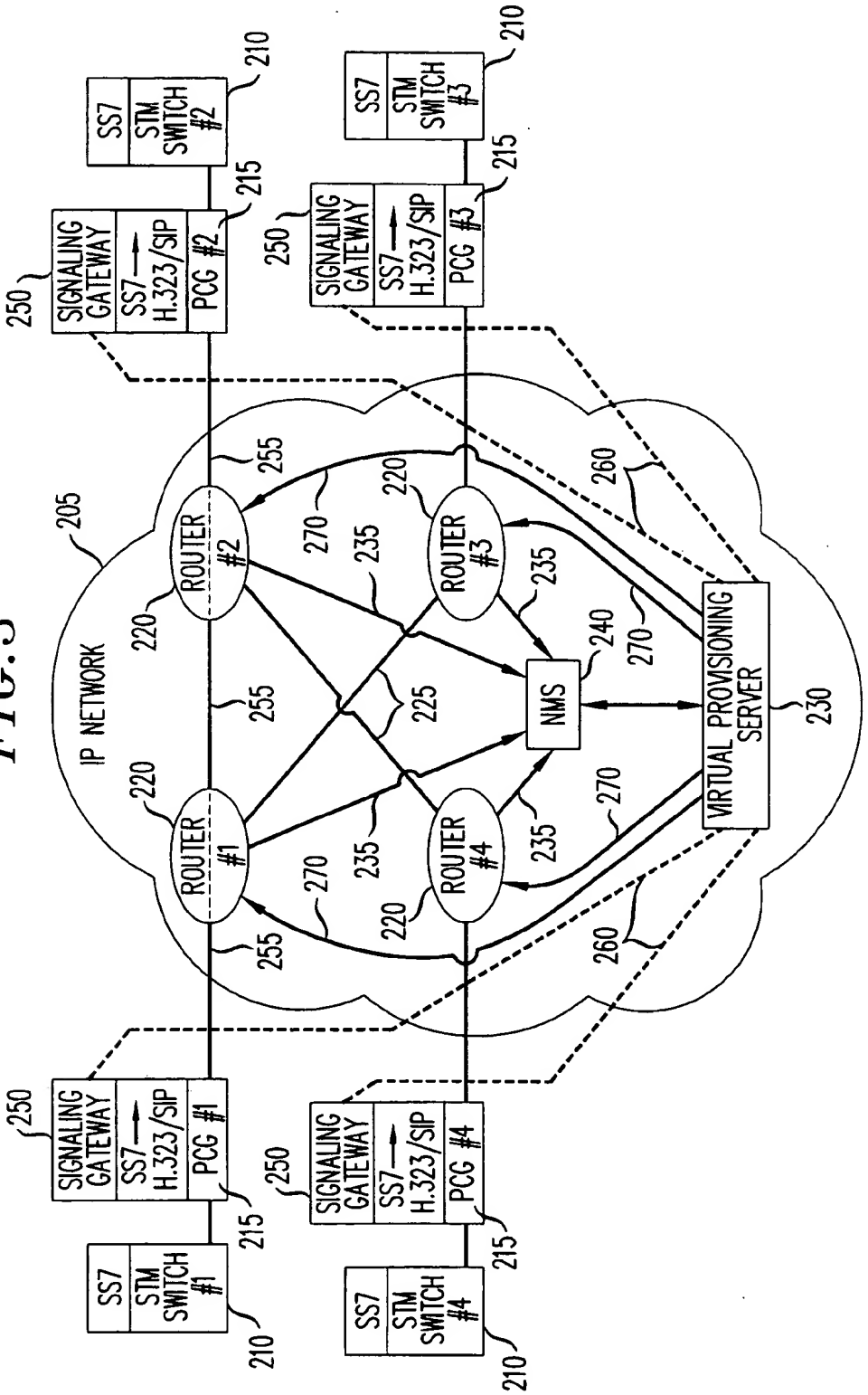

IP NETWORK
205
ROUTER #1
ROUTER #2
ROUTER #3
ROUTER #4
220
NMS
240
VIRTUAL PROVISIONING SERVER
230
SIGNALING GATEWAY
SS7 → H.323/SIP
PCG #1
PCG #2
PCG #3
PCG #4
250
215
SS7 STM SWITCH #1
SS7 STM SWITCH #2
SS7 STM SWITCH #3
SS7 STM SWITCH #4
210
225
235
255
260
270

FIG. 4

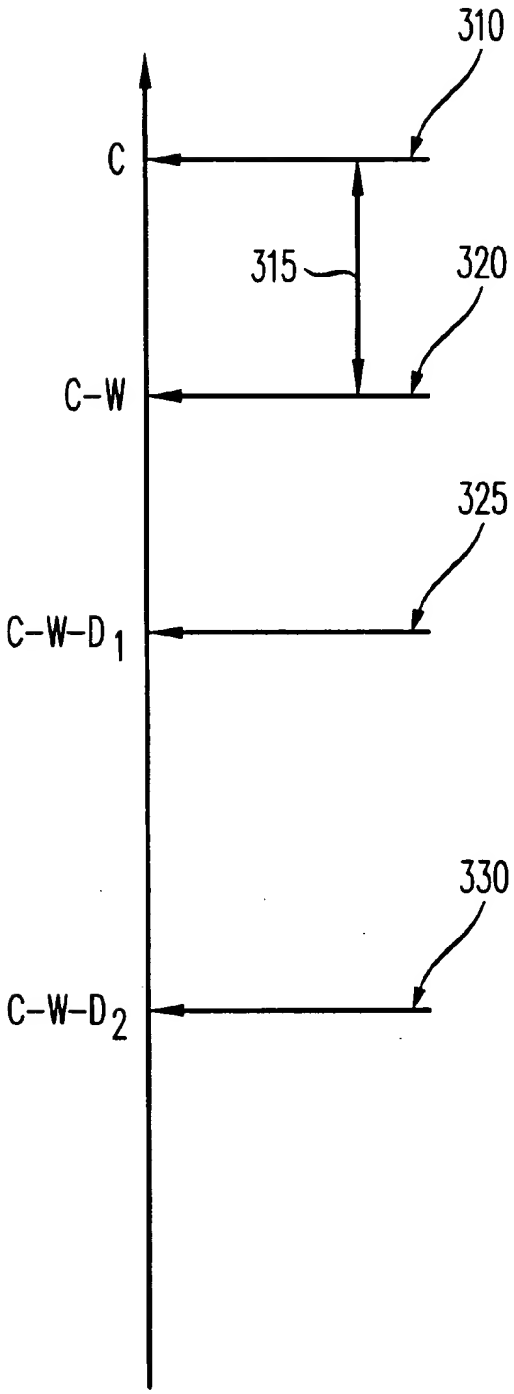
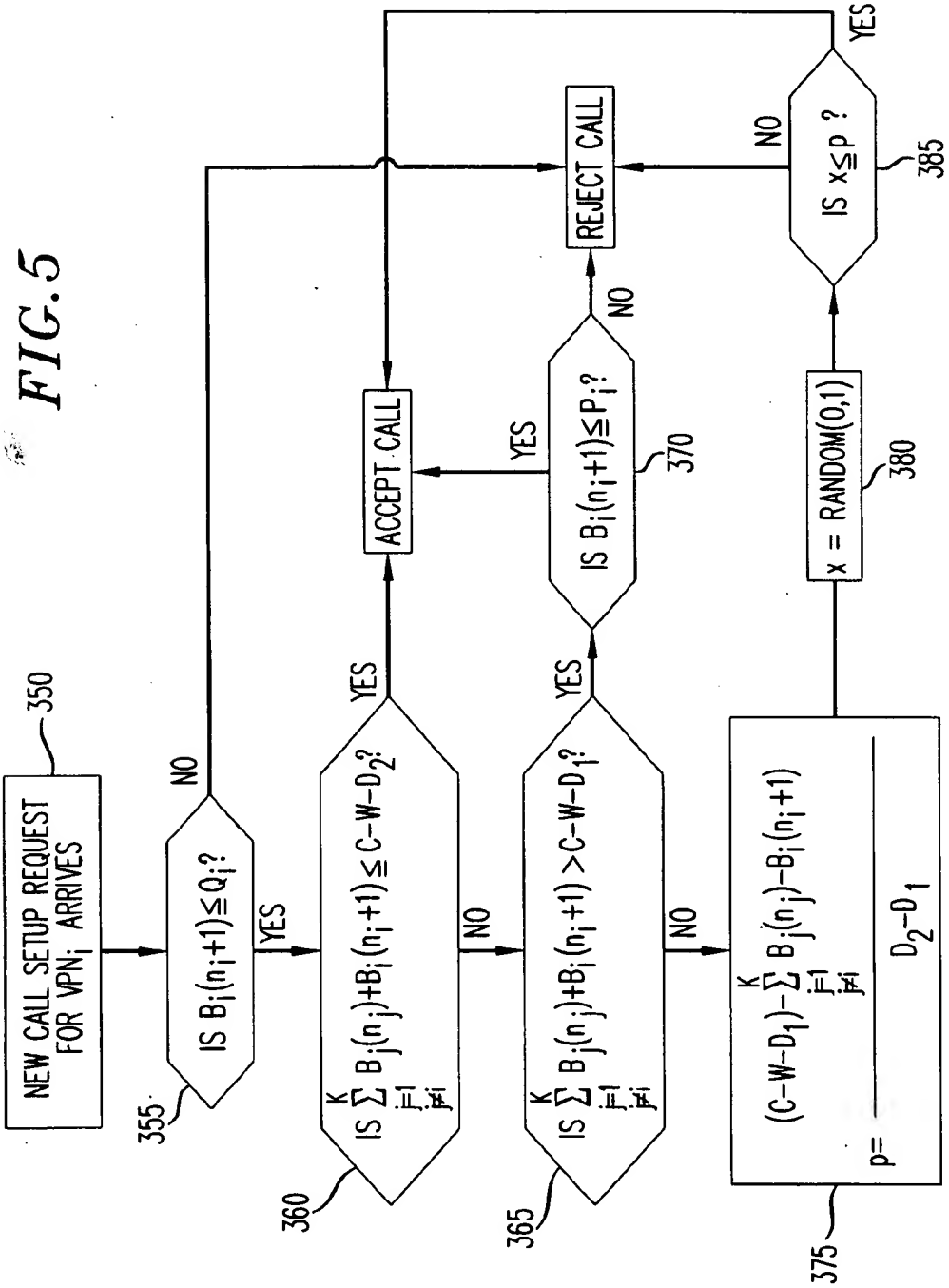


FIG. 5


NEW CALL SETUP REQUEST FOR VPN$_i$ ARRIVES
350
IS $B_i(n_i+1) \leq Q_i$?
355
NO
YES
IS $\sum_{j=1, j \neq i}^{K} B_j(n_j)+B_i(n_i+1) \leq C-W-D_2$?
360
YES
NO
IS $\sum_{j=1, j \neq i}^{K} B_j(n_j)+B_i(n_i+1) > C-W-D_1$?
365
YES
NO
$p = \frac{(C-W-D_1)-\sum_{j=1, j \neq i}^{K} B_j(n_j)-B_i(n_i+1)}{D_2-D_1}$
375
x = RANDOM(0,1)
380
IS $x \leq p$?
385
YES
NO
IS $B_i(n_i+1) \leq P_i$?
370
YES
NO
ACCEPT CALL
REJECT CALL

METHOD FOR PROVIDING QUALITY OF SERVICE FOR DELAY SENSITIVE TRAFFIC OVER IP NETWORKS

FIELD OF THE INVENTION

The present invention relates to the field of Internet Protocol (IP) networks, and more particularly to the transport of delay sensitive traffic over IP networks.

BACKGROUND OF THE INVENTION

A global network infrastructure for voice services, using a circuit-switching methodology, is supported by Public Switched Telephone and Private Branch Exchange networks. These networks utilize signaling to establish call connections and routing maps at network switches. The ability to signal during call connection set-up provides individual switches with the capability to reject call connection requests when that individual switch does not have the available bandwidth to support a new call connection. Since any switch in a connection path may reject a new call connection request based on available bandwidth limitations, switched voice networks are able to provide guaranteed Quality of Service to established connections. Quality of Service in switched voice networks is guaranteed because the governing precept is that it is preferable to block new call connection attempts rather than allow a new connected call to degrade the performance of established connected calls.

Explosive growth in Internet Protocol (IP) based Intranets and public Internet has generated a large network infrastructure of IP based routers. Recently, this large IP network infrastructure has begun to be utilized as a vehicle for real-time transmission of voice over the Internet, also known as Internet telephony. Each year, Internet telephony captures a greater share of the telephony market. However, unlike the case of switched voice service networks, routers contained within IP networks are not signaled. Since signaling between source, destination, and intermediate routers is not provided within IP networks, new calls can not be rejected at the IP routers, even if the routers are burdened beyond their respective bandwidth capacities. Therefore, real-time transmission over the Internet is subject to levels of delay and jitter not associated with Public Switched Telephone Networks and Private Branch Exchanges. Rather, transmission over the Internet and other IP networks is accomplished via a best effort transmission mode. Consequently, telephony over IP networks does not currently provide a Quality of Service guarantee for voice and other delay sensitive transmissions.

SUMMARY OF THE INVENTION

A quality of service guarantee for voice and other delay sensitive transmissions within an Internet Protocol (IP) network is provided by identifying the IP network path utilized for IP packet transmission between source and destination edge devices and virtually provisioning IP network path bandwidth for priority voice traffic. Priority for voice packets and admission control of new voice calls (and other delay sensitive traffic) based on the remaining available capacity over the IP network path guarantees that high priority voice (and other delay sensitive traffic) meet stringent delay requirements. A Virtual Provisioning Server is utilized to maintain bandwidth capacity data for each path segment within the IP network and to forward the bandwidth capacity data to a Signaling Gateway. The Signaling Gateway determines whether to accept or reject an additional delay sensitive traffic component based upon available bandwidth capacity for an IP network path. The Signaling Gateway then signals the originating source edge device as to its determination to accept or reject. Quality of Service guarantees concerning acceptable delay and jitter characteristics for real-time transmission over an IP network are therefore provided without the need to directly signal the individual IP routers over which an IP network path is established.

BRIEF DESCRIPTION OF THE DRAWINGS

A more complete understanding of the present invention may be obtained from consideration of the following description in conjunction with the drawings in which:

FIG. **1** is a diagram illustrating a voice over IP network between Packet Circuit Gateway edge devices and incorporating a Virtual Provisioning Server, the Virtual Provisioning Server communicating with a plurality of Signaling Gateways, in accordance with an exemplary embodiment of the present invention;

FIG. **2** is a diagram illustrating a voice over IP network between Packet Circuit Gateway edge devices and incorporating a Virtual Provisioning Server, the Virtual Provisioning Server communicating with a Signaling Gateway co-located with one Packet Circuit Gateway, and providing Signaling Gateway functionality to more than one Packet Circuit Gateway **215** within the network, in accordance with an exemplary embodiment of the present invention;

FIG. **3** is a diagram illustrating a voice over IP network between Packet Circuit Gateway edge devices and incorporating a Virtual Provisioning Server, the Virtual Provisioning Server further performing functions as a Virtual Private Network (VPN) Resource Manager, in accordance with an exemplary embodiment of the present invention;

FIG. **4** is a diagram illustrating the bandwidth allocation structure associated with an exemplary embodiment of the present invention; and

FIG. **5** is a flow diagram illustrating one exemplary embodiment of an algorithm for call admission control for a plurality of Virtual Private Networks sharing a link within a common network, in accordance with the present invention.

DETAILED DESCRIPTION

FIGS. **1**, **2**, and **3** are diagrams illustrating various embodiments for IP networks **205** between Packet Circuit Gateway edge devices **215** incorporating a Virtual Provisioning Server **230**, in accordance with the present invention. In FIG. **1**, the Virtual Provisioning Server **230** communicates with a Signaling Gateway **250** associated with each Packet Circuit Gateway edge device **215**. In FIG. **2**, the Virtual Provisioning Server **230** communicates with a Signaling Gateway **250** co-located with one Packet Circuit Gateway **215**, and providing Signaling Gateway functionality to more than one Packet Circuit Gateway **215** within the network. In FIG. **3**, the Virtual Provisioning Server **230** performs additional functions as a Virtual Private Network Resource Manager.

The present invention is described as being utilized within an environment wherein voice traffic originates and terminates on regular Public Switched Telephone Network circuit switches, such as Synchronous Transfer Mode switches **210**, and is carried over paths between routers within an IP network **205**. However, these circuit switches may also be implemented as simple access multiplexers or edge vehicles as would be apparent to those skilled in the art. It would also

be apparent to those skilled in the art that the present invention may be practiced with any IP datagram traffic (in addition to voice), although the present invention provides the greatest benefit for the transport of delay sensitive IP datagram traffic. Conversion from a circuit signal to IP format occurs at Packet Circuit Gateways (PCGs) **215**, which are also alternatively known as Service Access Concentrators (SACs) or Internet Telephone Gateways. In addition to conversion between circuit and IP formats, Packet Circuit Gateways **215** also provide voice compression/decompression, silence suppression/insertion, and other well known functions needed for specific applications.

Signaling Gateways **250** are utilized to provide the appropriate interface and interworking between signaling mechanisms and also to determine acceptance or rejection of a new call request originating from an associated Packet Circuit Gateway. Circuit networks, such as Public Switched Telephone Networks, typically use Signaling System 7 (SS7) to communicate requests for connection set-up and tear down. IP endpoints and intermediate routers use ITU-T H.323 or Session Initiation Protocol (SIP) for session management. Therefore, Signaling Gateways **250** provide a higher layer protocol utilized at the Packet Circuit Gateways **215** to facilitate conversions in signaling mechanisms between Public Switched Telephone Networks and IP networks **205**. It should be noted that a resident Signaling Gateway **250** is not required at each Packet Circuit Gateway Rather, the Signaling Gateway function may be implemented at a single location for all Packet Circuit Gateways with control signals transmitted to corresponding Packet Circuit Gateways from the single Signaling Gateway. For example, FIGS. **1** and **3** illustrate embodiments of the present invention wherein each Packet Circuit Gateway **215** maintains a resident Signaling Gateway **250**. However, FIG. **2** illustrates an embodiment of the present invention wherein only PCG #1 maintains a resident Signaling Gateway **250**. The Signaling Gateway functions are provided at PCG#2, PCG#3, and PCG#4 by transmission of appropriate control signals between the Signaling Gateway resident at PCG#1 and the remaining Packet Circuit Gateways. Transmission may be over the serviced IP network **205** within a TCP/IP session, an adjunct transmission medium, or any other well known means for data transport.

One unique feature of the present invention is provided by a Virtual Provisioning Server **230**. The Virtual Provisioning Server is utilized to provide the Signaling Gateways **250** with network bandwidth capability information, so that the Signaling Gateways are able to make a determination as to whether to accept or reject a new call request at an associated Packet Circuit Gateway **215**. The basis for admission/denial decisions for new calls is made in order to provide assurances that Quality of Service characteristics, such as delay, jitter, and loss of call connections, are maintained below a guaranteed threshold for established voice call connections.

The Virtual Provisioning Server **230** communicates the network bandwidth capability information to the Signaling Gateways **250** at least once at the commencement of network operation, and episodically whenever the underlying IP network is subject to changes to its link bandwidths due to link failures, new link establishment, addition of bandwidth to existing links, etc. A Network Management System (NMS) is typically associated with an IP network and its functions well known in the art. However, in association with the present invention the Network Management System performs the additional function of apprising the Virtual Provisioning Server of any changes to the link bandwidths as enunciated above.

FIGS. **1–3** illustrate a network path **255** for the transport of IP packets between PCG#1 and PCG#2. The path **255** is via intermediate components Router #1 and Router #2. Routers **220** are interconnected at the physical layer within the IP network **205** by a plurality of physical layer router transport segments **225**. It is over a plurality of these physical layer router transport segments **225** that the illustrated network path **255** is established. A network path **255** is comprised of a plurality of path links established over the plurality of physical layer router transport segments **225**. The Virtual Provisioning Server **230**, in cooperation with the Public Switched Telephone Network provisioning mechanism and admission control implemented by the Signaling Gateway **250**, provides for a quality guarantee to voice traffic while allowing the remaining capacity in the IP network to be used by other traffic utilizing the well known best effort mode. Similar provisioning can extend the service guarantee to multiple classes of traffic, for example—video conferencing.

Given that specific STM switches **210** are tied to corresponding Packet Circuit Gateways **215**, voice call transport capacity can be easily predicted using standard traffic engineering methods to determine the capacity needed between Packet Circuit Gateways **215**. Specific format variables, such as the type of compression method used, the silence suppression capability, etc., determine the network path bandwidth requirements between each pair of Packet Circuit Gateways **215**. The Virtual Provisioning Server **230** maintains and manages data corresponding to the transmission capacities of the IP network routers **220** and the physical layer router transport segments **225** between those routers **220**. The Virtual Provisioning Server is used, in accordance with the present invention, to determine the capacity requirements over each path between IP network routers **220** to meet the needed bandwidth requirements between Packet Circuit Gateways **215**. The capacity requirements over each network element, such as routers **220** and physical layer router transport segments **225** are virtually provisioned within available bandwidth capacity for delay sensitive traffic requirements. In accordance with the present invention, the bandwidth is considered virtually provisioned since the admission/denial of new connected calls is not controlled at each individual router **220**, but rather at the Packet Circuit Gateway edge devices **215**. Remaining bandwidth capacity over network elements is made available to delay insensitive packet transport only after the provisioning of bandwidth for delay sensitive voice frames or IP packets at the Packet Circuit Gateways **215** is performed. Alternatively, a provisioned minimum bandwidth capacity over each IP network path may be reserved for delay insensitive traffic, with the remaining bandwidth allocated for use by delay sensitive traffic. A Type-of-Service (TOS) field in the IP packet header is utilized to distinguish between delay sensitive and delay tolerant traffic types. Thus, voice packets may be given priority over data packets to ensure that delay and packet loss is in accordance with Quality of Service requirements.

If IP network routers **220** and physical layer router transport segments **225** utilized for a specific path **255** do not have the necessary bandwidth capacity to meet determined capacity requirements, the Virtual Provisioning Server **230** allocates portions of the bottleneck capacity to the pairs of Packet Circuit Gateways **215** competing for this capacity and messages the associated Signaling Gateway **250** of this allocation. The Virtual Provisioning Server **230** also calculates the need for added capacity within the IP network **205** to meet current and future bandwidth needs. By centrally

calculating and determining required network bandwidth provisioning and messaging the Signaling Gateways **205** within the IP network **205** of the bandwidth allocation, the Virtual Provisioning Server **230** determines the maximum number of voice calls that can be supported simultaneously between any pair of Packet Circuit Gateways **215**. Since Signaling Gateways **250** provide the signaling interworking between SS7 and H.323/SIP, they are also able to track the number of connected calls in progress between pairs of Packet Circuit Gateways **215**. As shown in the embodiment of the present invention illustrated in FIG. **2**, and as previously described, one Signaling Gateway **250** may be utilized to control more than one Packet Circuit Gateway **215** and may also be utilized to track the number of connected calls in progress between other network Packet Circuit Gateways **215** (PCG #**2**, PCG #**3**, and PCG #**4** in the instant embodiment as illustrated in FIG. **2**).

As previously described, the Virtual Provisioning Server **230** also exchanges data with a Network Management System (NMS) **240**. The Network Management System is a well known network controller used to maintain IP network **205** information pertaining to network element capacities, network bandwidth and capacity demand and growth data, link failures, etc. The Network Management System **240** is operable to exchange messages and signals with network routers **220** and to provide and maintain this network information via signaling channels **235**. However, the Network Management System **240** does not determine or control admission/denial decisions for new call connections at the Packet Circuit Gateways **215**. The Network Management System **240** provides the Virtual Provisioning Server **230** with information about the IP network **205** topology, capacities, failure events, etc. The Virtual Provisioning Server **230** uses this information to update its calculations and signals the Network Management System **240** if changes need to be implemented within the IP network, such as updating routing algorithm weights. Routing algorithm weights are used to determine the routing path for forwarding an IP packet. The use and implementation of such routing algorithm weights is well known in the art of IP networking. When needed capacities cannot be achieved temporarily due to failure events, the Virtual Provisioning Server **230** determines the maximum number of calls that can be supported on affected paths throughout the network and informs the associated Signaling Gateways **250**, thereby providing a mechanism to throttle the number of connected calls at the various network Packet Circuit Gateway edge devices **215**.

Although the instant embodiment of the present invention is described in the context of connectivity between PSTN switches and Signaling Gateways **250** to manage signaling conversion and admission control, it may also be used to support telephony between PCs and telephony between a PC and a phone via a PSTN switch. In order to guarantee connection quality for these connections, it is important to provide messaging from the Virtual Provisioning Server **230** to the Signaling Gateway **250**, thus informing the Signaling Gateway about the call capacities for PCG-to-PCG paths for a minimum of telephony traffic originating from PSTN and PCs. In addition, since a network operator may not control the coding rate in this case (i.e.—when calls originate from PCs), a traffic policing function is utilized at the PCG to monitor compliance with the traffic assumptions used in call set-up signaling.

Voice calls originating from a PC may be assigned lower priority as compared to those originating from a PSTN. Doing so allows the Signaling Gateway **250** to reject PC originated calls based on a lower bandwidth utilization, and rejects the PSTN originated calls at a higher threshold. Therefore, the Signaling Gateway **250** can guarantee call connection quality for voice and other Quality of Service sensitive services by enforcing call admission control at the Packet Circuit Gateways **230** and preferentially awarding priority for PSTN originated voice services over other services. In addition, a service provider may provide a plurality of critical service guarantees to customers and similarly, multiple customers may desire similar critical service guarantees over common paths within an IP network **205**. One such example is presented within the context of Virtual Private Networks for voice traffic, wherein a network provider provides wide area services to interconnect corporate users in different locations. The ability to provide multiple Virtual Private Networks along with public service over a common infrastructure is attractive to both the service provider and corporate customers. One critical benefit of providing a Virtual Private Network is that the service provider is able to deliver secure access to the user. A second benefit is the ability to provide a Quality of Service guarantee comparable to that on leased private lines between customer premises switches (e.g., PBXs).

Virtual Private Network customers negotiate bandwidth and service quality guarantees from a wide area network operator or service provider. The network operator guarantees this negotiated service level to all Virtual Private Network customers by utilizing the common infrastructure to achieve multiplexing gain. Capabilities available in currently available routers **220** allow the Virtual Provisioning Server **230** to provide these guaranteed services. For example, routers are available which are capable of identifying flows based on the port, source, and destination identifiers, and which categorize group flows into classes and/or super classes according to the level of service and bandwidth guarantees negotiated. These routers are also operable to allocate and manage minimum and maximum bandwidth for each class, super class, etc. Incorporation of buffer and queue management at the routers provides distinction and differentiation of priority treatment among flow classes and super classes. Additionally, statistical multiplexing may be provided for flows within a class and/or among classes within a super class. A system of Weighted Fair Queuing (WFQ) service provides for management of flow, class, and super class bandwidths. If one of the classes or super classes exceeds a negotiated bandwidth allocation, superior service quality may still be provided if the other negotiated classes or super classes are not completely utilizing their allocated bandwidth. Therefore, only the Quality of Service provided to classes or super classes exceeding their negotiated allocation of bandwidth are affected.

Referring to FIG. **3**, the Virtual Provisioning Server **230** is utilized as a Virtual Private Network Resource Manager. The Virtual Private Network Resource Manager utilizes optimizing algorithms to (i) partition bandwidth between Virtual Private Networks and within Virtual Private Networks if the customer desires a further subclassification of services and (ii) control flow routing within the network. If the network routers **220** utilized have flow partitioning capability, but do not have flexible routing capability, then flow routes are fixed through the IP network **205** and capacities are partitioned in the network by the Virtual Private Network Resource Manager based upon the negotiated Virtual Private Network contract. The Virtual Provisioning Server **230**, functioning as a Virtual Private Network Resource Manager, sends this partitioning information to individual routers **220** within the network **205** so that the

network routers **220** are able to set algorithm weights, minimum bandwidth, maximum bandwidth, buffer thresholds, etc. Communication between the Virtual Private Network Resource Manager is illustrated over a VPN signaling path **270** between the Virtual Provisioning Server **230** and individual routers, in accordance with FIG. **3**. The illustrated VPN signaling path **270** is merely illustrative, and any number of other means for signaling routers **220** would also be apparent to those skilled in the art, including communicating through the Network Management System **240**. Once partitioning information is received at network routers **220** and partitioning is accomplished, each Virtual Private Network is established with its allocated minimum bandwidth.

Referring again to FIGS. **1–3**, Virtual Private Networks for voice may also be supported using PSTN switches or multiplexers as access vehicles (STM switches **210** in the instant example) and utilizing the IP network **205** as backbone, as was previously described. Advantageously, the instant embodiment for establishing Virtual Private Networks for voice is achieved using network routers **220** with simple priority mechanisms. That is, signaling is not required between the Virtual Provisioning Server **230** and network routers **220** to establish and maintain the Virtual Private Networks. Rather, the Virtual Provisioning Server **230** uses aggregate capacity needed between a pair of gateways to perform virtual provisioning. The Packet Circuit Gateways **215**, in conjunction with the Signaling Gateways **250**, are utilized to control the acceptance or rejection of new calls from each Virtual Private Network customer utilizing an acceptance/rejection algorithm residing in the Virtual Provisioning Server **230**.

FIGS. **4** and **5** illustrate and define an exemplary algorithm for performance of the acceptance or rejection of new calls over a Virtual Private Network established between Packet Circuit Gateways **215**, in accordance with the present invention. In conjunction with the accompanying description, the following definitions apply:

C=The total link bandwidth **310**,

W=The minimum bandwidth always available for combined traffic supported using Available Bit Rate (ABR) or best effort data service **315**,

C–W=The total bandwidth available for call admission control purposes **320**,

C–W–D_1=An upper threshold for call admission control purpose **325**,

C–W–D_2=A lower threshold for call admission control purpose **330**,

$B_i(n_i)$=Bandwidth needed to support n_i connections for VPN_i with a specified Quality of Service,

P_i=Minimum bandwidth contracted for VPN_i,

Q_i=Maximum bandwidth contracted for VPN_i, and

K=Number of Virtual Private Networks with Quality of Service guarantees sharing the link in consideration.

When a new call set-up request for VPN_i arrives at the Signaling Gateway **250**, then the exemplary algorithm associated with FIG. **5** is performed to determine whether to accept or reject the new call, in accordance with step **350**. The bandwidth utilized by K Virtual Private Networks (VPN_i; i=1,2,3, . . . K) is monitored at the Signaling Gateway **250**. Referring to step **355**, when the VPN_i bandwidth necessary to support an additional call exceeds the maximum bandwidth allocation (Q_i), the requested new call is rejected. However, when the VPN_i bandwidth necessary to support an additional call does not exceed the maximum bandwidth allocation (Q_i), then step **360** is performed. In accordance with step **360**, if the VPN_i bandwidth usage would be between the range of zero to (C–W–D_2) after connecting the new call, then the new call is accepted. However, if VPN_i bandwidth usage would be greater than (C–W–D_2) after connecting the new call, then step **365** is performed. In accordance with step **365**, if VPN_i bandwidth usage would be between the range from (C–W–D_1) to (C–V), a new call set-up request for VPN_i is accepted only if the bandwidth usage by VPN_i has not exceeded its minimum allocation, P_i, otherwise the call is rejected, in accordance with step **370**. If however, the VPN_i bandwidth usage is between the range of (C–W–D_2) to (C–W–D_1), a new call set-up request for VPN_i is accepted or rejected probabilistically based on a sliding scale algorithm in accordance with step **375**. Let q=(1–p) denote the ratio of bandwidth usage in excess of (C–W–D_2) over (D_2–D_1). A random number x is generated at the Signaling Gateway **250** to support the probabilistically based algorithm, in accordance with step **380**. If the value of x is less than or equal to probability p, then the new call is accepted, in accordance with step **385**. For a call that traverses multiple links between its source and destination PCGs, the algorithm of FIG. **4** and FIG. **5** is repeated for each path link used to establish the call. The call is connected between the source and destination PCGs only if the algorithm yields a positive determination (to accept the call) for each link in the path.

During implementation of the exemplary algorithm of FIG. **5**, the bandwidth utilization data, $B_i(n_i)$, as a function of the number, n_i, for calls over VPN_i is utilized. If the calls or connections are constant bit rate, then $B_i(n_i)$ is a simple linear function of n_i. However, if the calls or connections are variable bit rate by nature or by design, for example—voice with silence elimination, on/off data sources, etc., then $B_i(n_i)$ is typically a non-linear function of n_i. The non-linear nature of $B_i(n_i)$ is due to the statistical multiplexing of randomly varying variable bit rate sources, as is well known in the art. For example, the specific nature of a $B_i(n_i)$ function, in the context of packet voice multiplexing, is detailed in a publication by K. Sriram and Y. T. Wang entitled "Voice Over ATM Using AAL2 and Bit Dropping: Performance and Call Admission Control," Proceedings of the IEEE ATM Workshop, May 1998, pp. 215–224, which is incorporated herein by reference.

Prior reference to the Virtual Provisioning Server (VPS) is described in the context of an IP network which includes multiple interconnected Open Shortest Path First (OSPF) domains. The present invention may also be implemented within an IP network comprised of multiple interconnected administrative areas, wherein each administrative area is comprised of multiple OSPF domains. Typically, each administrative area is an IP network belonging to an individual internet service provider or carrier, although such a configuration is not required. Such an embodiment of the present invention may be implemented with each administrative area having one gateway VPS. Each respective VPS may be co-located with the gateway router for that respective administrative area, although co-location is not a required aspect of the embodiment. Each pair of respective gateway VPSs determines the capacity requirements between their respective gateway routers. Further, each gateway VPS provides the necessary bandwidth capacity information between pairs of neighboring administrative areas to the VPSs located in each of the OSPF domains within its administrative area. Thus, the signaling gateways anywhere in the larger IP network are adequately provided with the necessary information for admission/denial of calls,

including those that originate in one administrative area and terminate in another.

Numerous modifications and alternative embodiments of the invention will be apparent to those skilled in the art in view of the foregoing description. For example, although the present invention has been described in the context of a single Virtual Provisioning Server utilized to service an entire IP network and control all Signaling Gateways within that network, it is also equally applicable for an embodiment of the present invention operable for multi-domain operation. That is, for those instances when call routing is made from a first telephony gateway source connected to a first IP domain and the destination is a second telephony gateway connected through another IP domain, the call processing involves intra-domain routing to the gateway router in the first domain, routing among gateway routers in intervening domains, and intra-domain routing from the gateway router to the telephony gateway in the last domain. Protocols such as Open Shortest Path First (OSPF) determine routing in a domain while a Border Gateway Protocol (BGP) is used for inter-domain routing between gateway domains. In such an embodiment of the present invention, a plurality of Virtual Provisioning Servers are utilized, one for each IP domain. Each Virtual Provisioning Server manages the virtual provisioning of routers within its respective domain, including Gateway Border Routers. Additionally, each pair of interfacing Virtual Provisioning Servers determines the capacity requirements between their respective pair of interfacing Gateway Border Routers. As was true for the single domain embodiment of the present invention, admission/denial control at the originating and terminating Packet Circuit Gateways is enabled without signaling the incorporated routers directly. In the multi-domain embodiment, this capability is attributable to shared knowledge of intra-domain and inter-domain routing protocols among the interfaced Virtual Provisioning Servers and also due to the static nature of router algorithm weights.

Additionally, the previous description is applicable for embodiments of the present invention in which service guarantees are provided without adding signaling mechanisms between routers and the associated Virtual Provisioning Server. However, the present invention would be equally applicable for those instances in which the Virtual Provisioning Server is operable to directly signal the network routers; although such an embodiment would be more accurately described as having a Server in which the provisioning is more real than virtual (since the provisioning is controlled at the routers instead of at the corresponding originating and terminating gateways). This alternative embodiment utilizes state exchange protocols in Open Shortest Path First (OSPF) and Border Gateway Protocol (BGP), which are extended to provide dynamic topology and capacity information.

The present invention may also be used in evolving IP networks in which the well-known Multi-Protocol Label Switching (MPLS) is utilized at the network IP routers. In an MPLS based IP network, the Virtual Provisioning Server maintains a knowledge base of possible multiple paths between source-destination pairs of Packet Circuit Gateway edge devices. The Signaling Gateways receive information from the Virtual Provisioning Server about alternative paths and associated capacities between PCG pairs, and admits a new voice call request if capacity is available over any of the available paths, otherwise, the call request is rejected.

Accordingly, this description is to be construed as illustrative only and is for the purpose of teaching those skilled in the art the best mode of carrying out the invention and is not intended to illustrate all possible forms thereof. It is also understood that the words used are words of description, rather than limitation, and that details of the structure may be varied substantially without departing from the spirit of the invention and the exclusive use of all modifications which come within the scope of the appended claims are reserved.

What is claimed is:

1. A method for providing a Quality of Service guarantee for delay sensitive traffic conveyed over a path within an Internet Protocol (IP) network having a virtual provisioning server, a source edge device providing an interface for launching said delay sensitive traffic within said IP network, said method comprising the steps of:

receiving, at a signaling gateway, a value representing a bandwidth capacity for said path;

receiving at said signaling gateway, a request to establish an additional delay sensitive traffic component over said path;

comparing, at said signaling gateway, said value representing said bandwidth capacity for said path with a total bandwidth needed if said additional delay sensitive traffic component is established over said path;

identifying, at said signaling gateway, at least one of a plurality of paths within said IP network as having a most limiting available bandwidth capacity, wherein the identified path has sufficient available bandwidth capacity to handle said additional delay sensitive traffic component; and

limiting said quantity of said delay sensitive traffic launched from said source edge device to less than or equal to said most limiting available bandwidth capacity.

2. The method in accordance with claim **1** wherein said value representing said bandwidth capacity for said path is transmitted from said virtual provisioning server to said signaling gateway.

3. The method in accordance with claim **1** wherein said request to establish said additional delay sensitive traffic component over said path is conveyed from said source edge device.

4. The method in accordance with claim **3** wherein said source edge device is a packet circuit gateway.

5. The method in accordance with claim **1** further comprising the step of conveying said signal denying said request to establish said additional delay sensitive traffic component from said signaling gateway to said source edge device.

6. The method in accordance with claim **1** further comprising the steps of:

generating, at said signaling gateway, a signal authorizing said request to establish said additional delay sensitive traffic component if said total bandwidth needed is less than or equal to said value representing said bandwidth capacity for said path; and

conveying said signal authorizing said request to establish said additional delay sensitive traffic component from said signaling gateway to said source edge device.

7. The method in accordance with claim **1**, further comprising:

generating, at said signaling gateway, a signal denying said request to establish said additional delay sensitive traffic component if said total bandwidth needed is greater than said value representing said bandwidth capacity for said path.

8. A method for providing a Quality of Service guarantee for real-time voice transmission traffic conveyed between a

source Packet Circuit Gateway and a destination Packet Circuit Gateway over an Internet Protocol (IP) network having a plurality of routers, said source Packet Circuit Gateway providing an interface for launching said real-time voice transmission traffic within said IP network over an IP network path, said method comprising the steps of:

partitioning, from a bandwidth capacity associated with said IP network path, a first provisioned bandwidth capacity for a first Virtual Private Network (VPN), said VPN contracted for said real-time voice transmission traffic conveyed between said source Packet Circuit Gateway and said destination Packet Circuit Gateway;

maintaining, at a Signaling Gateway, a value representing said first provisioned bandwidth capacity for said first VPN;

receiving, at said Signaling Gateway, a request from said source Packet Circuit Gateway to establish a new call connection with said destination Packet Circuit Gateway over said first VPN, in addition to a plurality of presently established call connections;

comparing, at said Signaling Gateway, said value representing said first provisioned bandwidth capacity for said first VPN with a required first VPN bandwidth capacity should said new call connection be established; and

transmitting, from said Signaling Gateway, a signal denying said request to establish said new call connection if said required first VPN bandwidth capacity should said new call connection be established is greater than said value representing said first provisioned bandwidth capacity for said first VPN.

9. The method in accordance with claim **8** further comprising the step of:

transmitting, from said Signaling Gateway, a signal authorizing said request to establish said new call connection if said required first VPN bandwidth capacity should said new call connection be established is less than or equal to said value representing said first provisioned bandwidth capacity for said first VPN.

10. The method in accordance with claim **8** wherein a Virtual Provisioning Server is utilized to provide said Signaling Gateway with said value representing said first provisioned bandwidth capacity for said first VPN.

11. The method in accordance with claim **10** wherein said Virtual Provisioning Server is adapted to maintain a plurality of Virtual Private Networks over said IP network path.

12. The method in accordance with claim **8** wherein said Quality of Service guarantee is established by maintaining delay of said real-time voice transmission traffic conveyed between said source Packet Circuit Gateway and said destination Packet Circuit Gateway below a guaranteed threshold value.

13. The method in accordance with claim **8** wherein said Quality of Service guarantee is established by maintaining jitter of said real-time voice transmission traffic conveyed between said source Packet Circuit Gateway and said destination Packet Circuit Gateway below a guaranteed threshold value.

14. The method in accordance with claim **8** wherein a circuit network switch is utilized to supply and accept said plurality of presently established call connections and said new call connection from said source Packet Circuit Gateway.

15. The method in accordance with claim **14** wherein said circuit network switch is a Synchronous Transfer Mode (STM) switch.

16. The method in accordance with claim **8** wherein at least one of said plurality of routers is operable to support Multi-Protocol Label Switching.

17. The method in accordance with claim **10** wherein a plurality of Multi-Protocol Label Switching (MPLS) routers is utilized to establish a plurality of paths between said source Packet Circuit Gateway and said destination Packet Circuit Gateway.

18. The method in accordance with claim **17** wherein said Virtual Provisioning Server is further operable to provide said Signaling Gateway with a plurality of values representing bandwidth capacities for each of said plurality of paths between said source Packet Circuit Gateway and said destination Packet Circuit Gateway.

19. The method in accordance with claim **10** wherein a plurality of Virtual Provisioning Servers are utilized to service a corresponding plurality of Open Shortest Path First domains.

20. The method in accordance with claim **10** wherein a plurality of Virtual Provisioning Servers are utilized to service a corresponding plurality of multiple administrative areas.

* * * * *

US006515964B1

(12) **United States Patent**
Cheung et al.

(10) **Patent No.: US 6,515,964 B1**
(45) **Date of Patent: Feb. 4, 2003**

(54) **METHOD AND APPARATUS FOR DYNAMICALLY CONTROLLING THE ADMISSION OF CALLS TO A NETWORK**

(75) Inventors: **Hay Yeung Cheung**, Holmdel, NJ (US); **Louise E. Hosseini-Nasab**, Holmdel, NJ (US); **Daniel J. Yantro, Jr.**, Middletown, NJ (US)

(73) Assignee: **AT&T Corp.**, New York, NY (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/303,305**

(22) Filed: **Apr. 30, 1999**

Related U.S. Application Data

(60) Provisional application No. 60/114,150, filed on Dec. 29, 1998.

(51) **Int. Cl.**[7] **H04L 12/56**
(52) **U.S. Cl.** **370/230**; 370/252; 370/352
(58) **Field of Search** 370/230, 238, 370/352, 356, 252, 253

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,193,151	A *	3/1993	Jain	370/230
5,400,329	A *	3/1995	Tokura et al.	370/232
5,732,078	A *	3/1998	Arango	370/355
5,796,719	A *	8/1998	Peris et al.	370/230
6,064,653	A *	5/2000	Farris	370/352
6,192,031	B1 *	2/2001	Reeder et al.	370/230
6,222,824	B1 *	4/2001	Marin et al.	370/230
6,226,266	B1 *	5/2001	Galand et al.	370/235
6,282,192	B1 *	8/2001	Murphy et al.	370/352
6,377,573	B1 *	4/2002	Shaffer et al.	370/356

OTHER PUBLICATIONS

Gordon, Shykeh, H.323: The Multimedia Communications Standard Moves From Consensus to Compliance, CTI Developer, 2(2):108–113.

Blank Michelle, H.323 Gatekeepers: Essential Software for IP Telephony and Multimedia Conferencing, CTI Developer, pp. 94–98, Feb. 1998.

Shenker, et al., RFC 2212 Specification of Guaranteed Quality of Service, pp. 1–20, Sep. 1997.

International Telecommunication Union, H.323, Series H: Audiovisual and Multimedia Systems: Infrastructure of audiovisual services—Systems and terminal equipment for audiovisual services, Visual telephone systems and equipment for local area networks which provide a non–guaranteed quality of service, 79 pp., Nov. 1996.

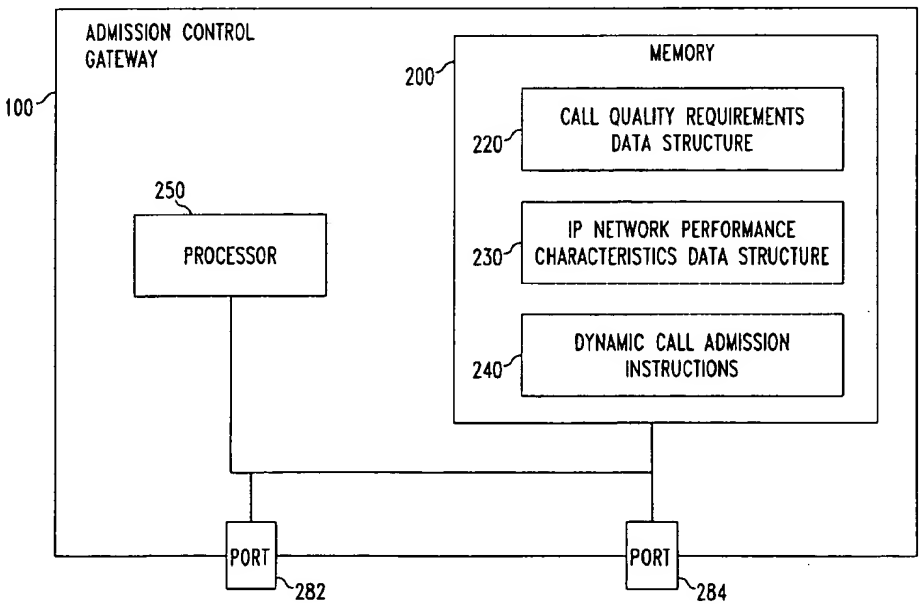
* cited by examiner

Primary Examiner—Chau Nguyen
Assistant Examiner—Keith M. George

(57) **ABSTRACT**

A network call admission control system receives a call and determines a call characteristic requirement and a network characteristic parameter. The call is admitted to the network based in part on whether the call characteristic requirement is satisfied by the network characteristic parameter. As a result, a communications service provider can provide a high quality of service for completed calls or charge a discounted rate for completed calls not meeting a certain quality of service.

3 Claims, 6 Drawing Sheets



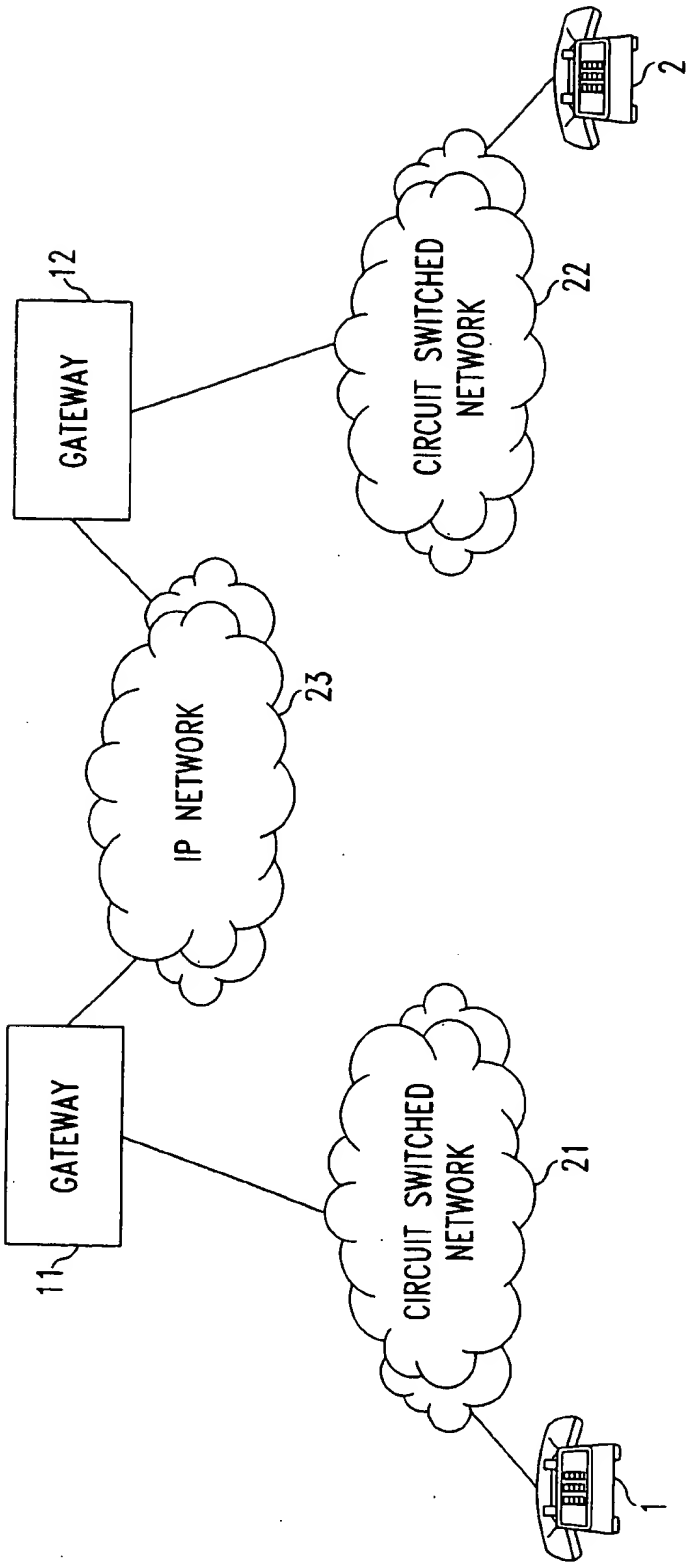


FIG. 1
PRIOR ART

FIG. 2

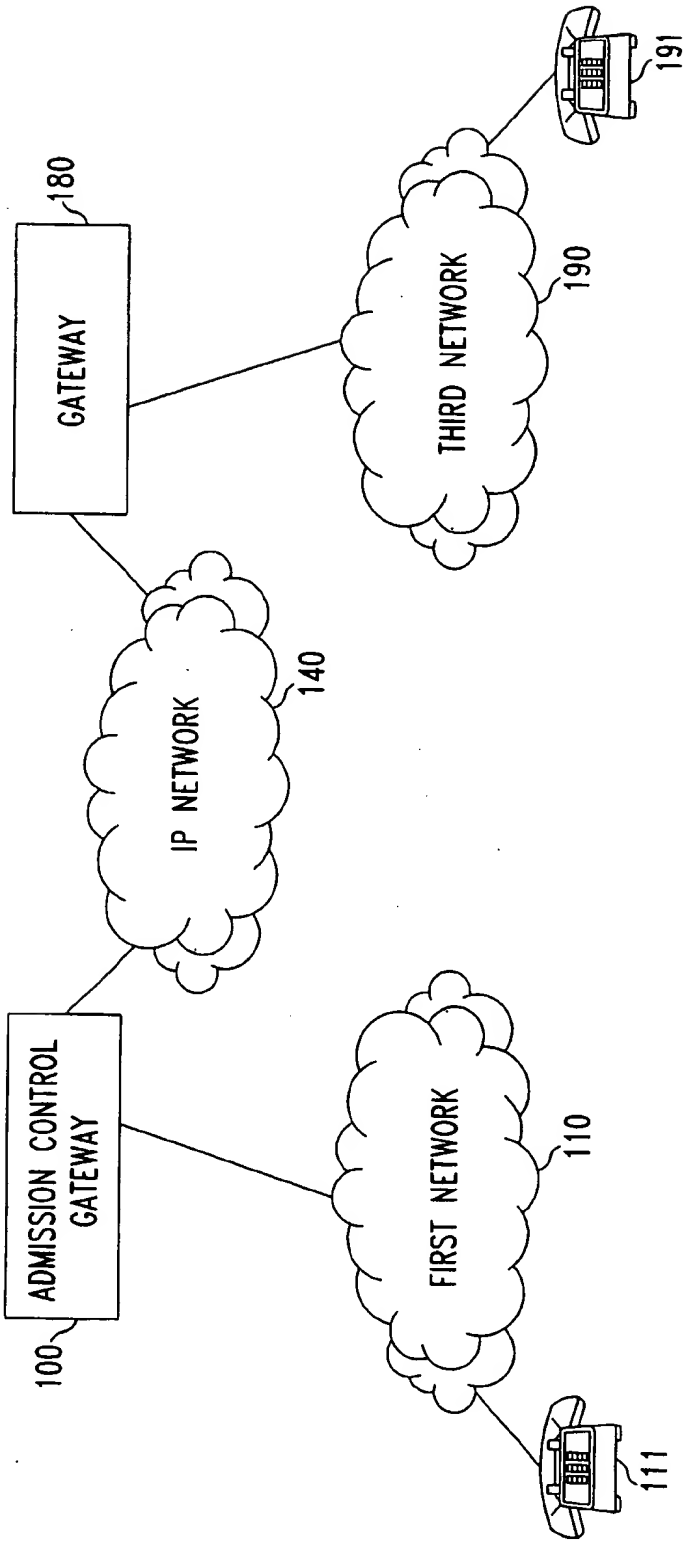
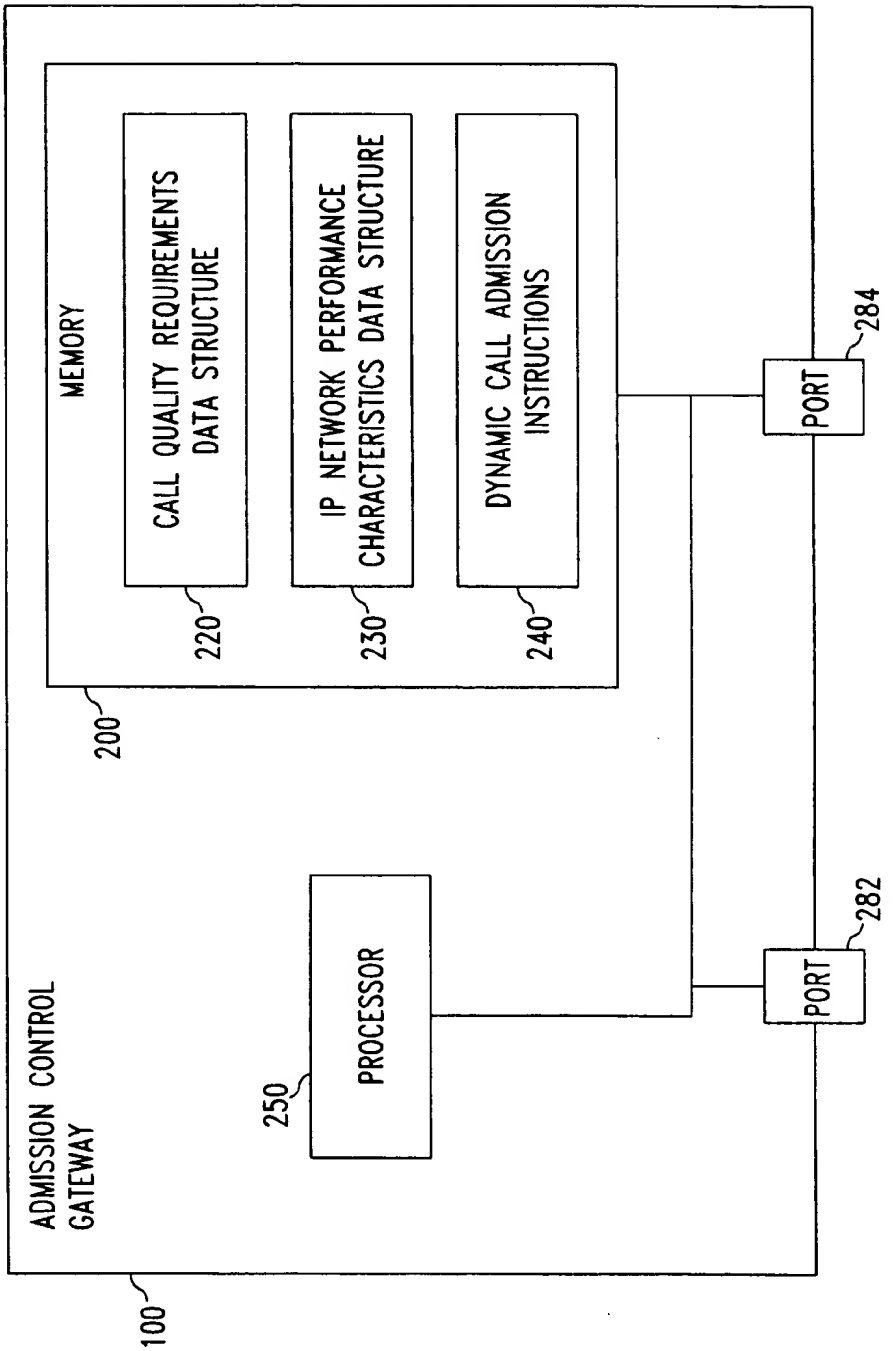

191
180
GATEWAY
THIRD NETWORK
190
IP NETWORK
140
ADMISSION CONTROL GATEWAY
FIRST NETWORK
110
100
111

FIG. 3


ADMISSION CONTROL GATEWAY
100
MEMORY
200
CALL QUALITY REQUIREMENTS DATA STRUCTURE
220
IP NETWORK PERFORMANCE CHARACTERISTICS DATA STRUCTURE
230
DYNAMIC CALL ADMISSION INSTRUCTIONS
240
PROCESSOR
250
PORT
282
PORT
284

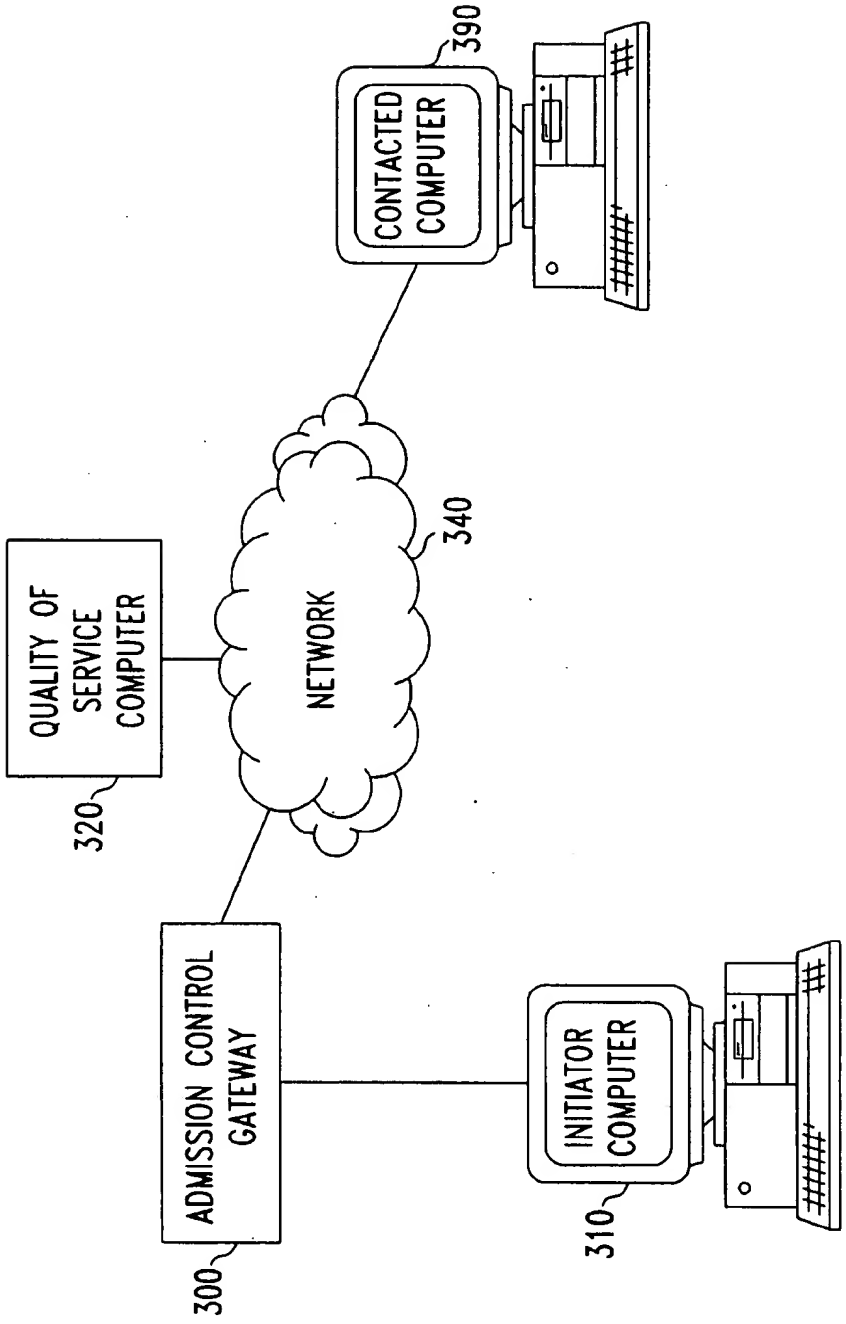
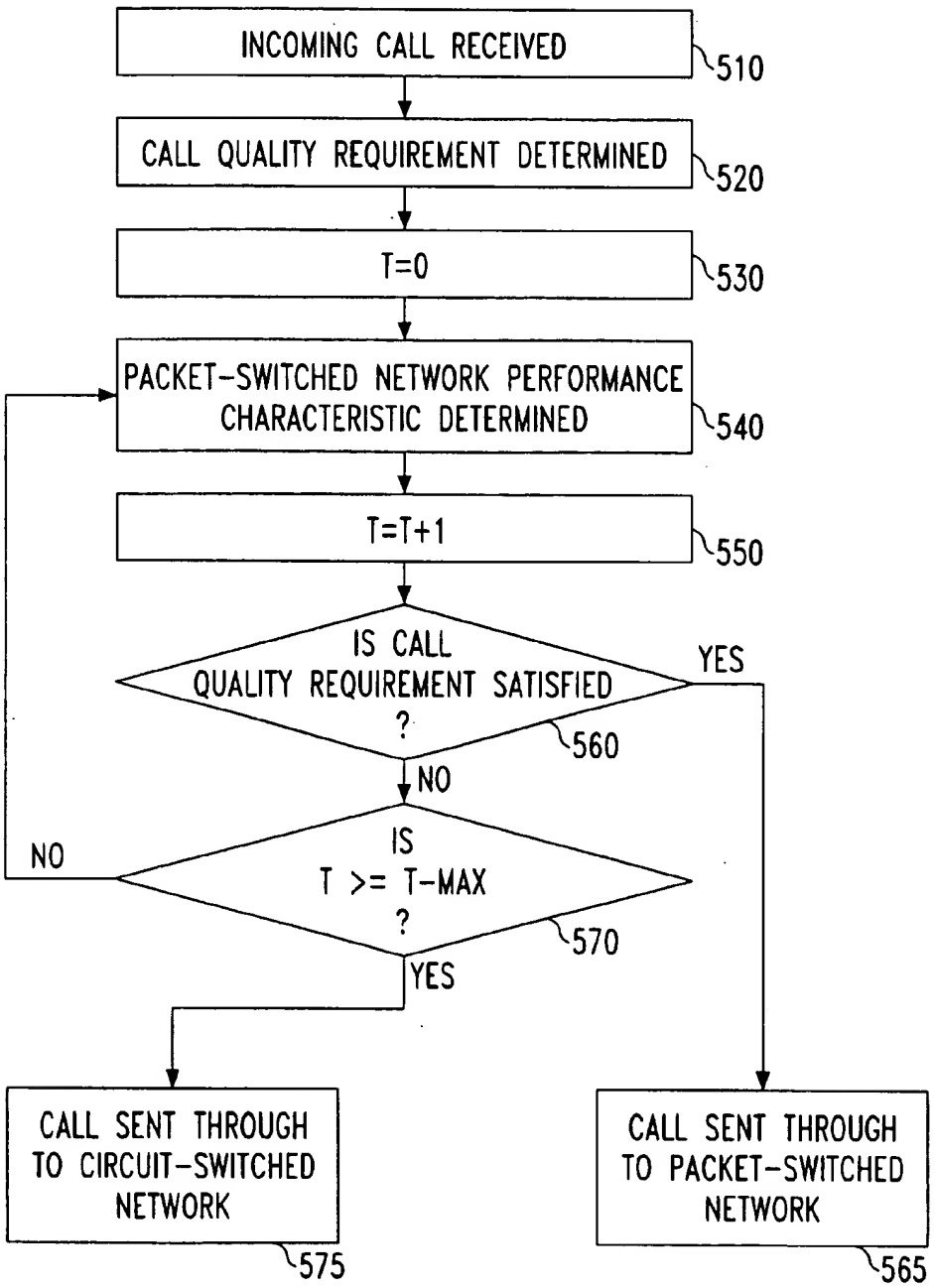

CONTACTED COMPUTER
390
NETWORK
340
QUALITY OF SERVICE COMPUTER
320
ADMISSION CONTROL GATEWAY
300
INITIATOR COMPUTER
310

FIG. 4

FIG. 5


INCOMING CALL RECEIVED
510
CALL QUALITY REQUIREMENT DETERMINED
520
T=0
530
PACKET-SWITCHED NETWORK PERFORMANCE CHARACTERISTIC DETERMINED
540
T=T+1
550
IS CALL QUALITY REQUIREMENT SATISFIED ?
560
YES
NO
IS T >= T-MAX ?
570
NO
YES
CALL SENT THROUGH TO CIRCUIT-SWITCHED NETWORK
575
CALL SENT THROUGH TO PACKET-SWITCHED NETWORK
565

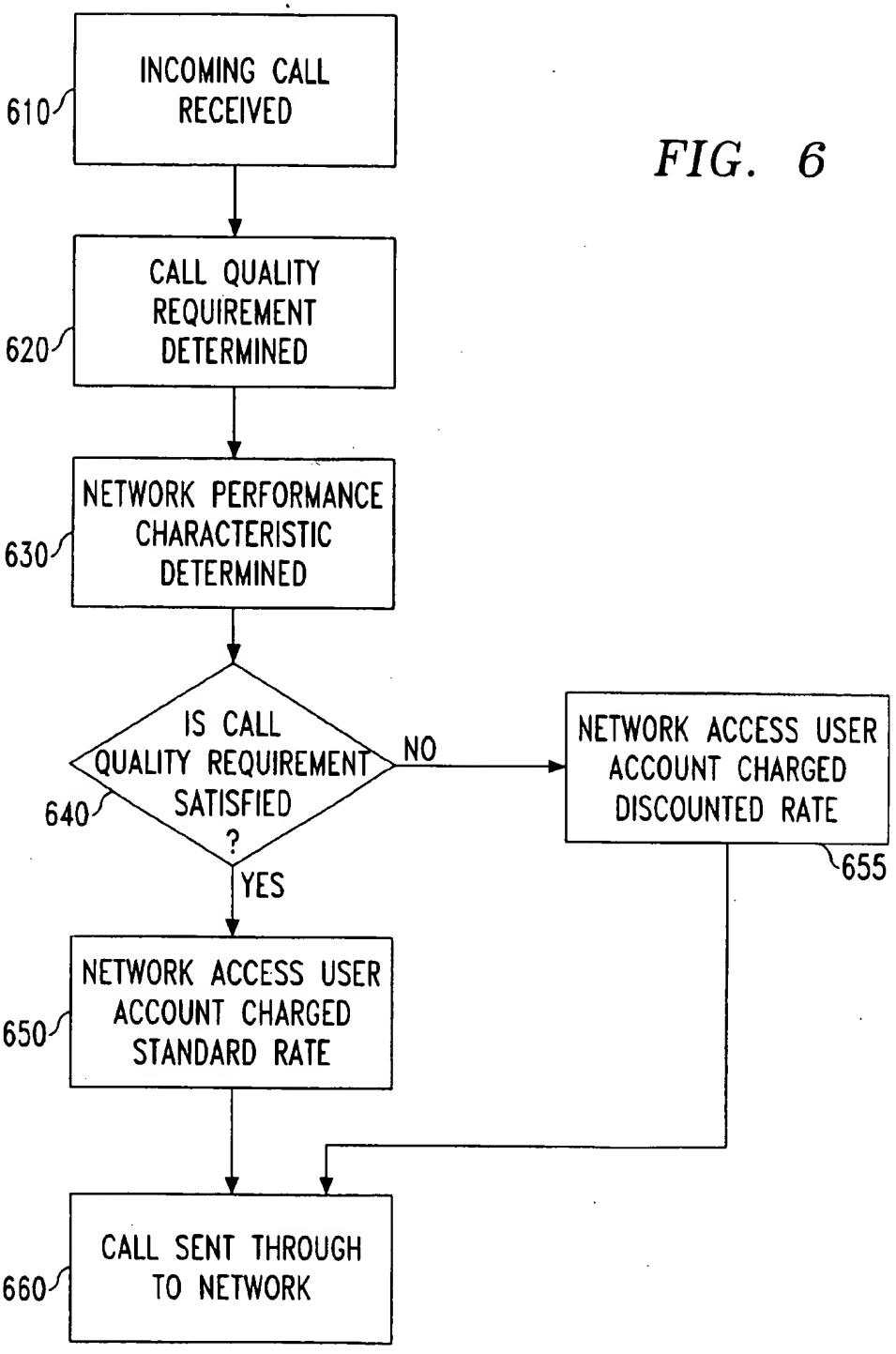

INCOMING CALL RECEIVED
610
CALL QUALITY REQUIREMENT DETERMINED
620
NETWORK PERFORMANCE CHARACTERISTIC DETERMINED
630
IS CALL QUALITY REQUIREMENT SATISFIED ?
640
NO
NETWORK ACCESS USER ACCOUNT CHARGED DISCOUNTED RATE
655
YES
NETWORK ACCESS USER ACCOUNT CHARGED STANDARD RATE
650
CALL SENT THROUGH TO NETWORK
660

FIG. 6

METHOD AND APPARATUS FOR DYNAMICALLY CONTROLLING THE ADMISSION OF CALLS TO A NETWORK

CROSS REFERENCE TO RELATED APPLICATION

The present application claims the benefit of U.S. provisional patent application Ser. No. 60/114,150 entitled "Method and Apparatus for Dynamically Controlling the Admission of Calls to a Network" to Daniel J. Yaniro, Louise E. Brown, and Hay Yeung Cheung and filed on Dec. 29, 1998.

FIELD OF THE INVENTION

The invention relates to network call admission. More particularly, the invention relates to a method and apparatus for dynamically controlling the admission of calls to a network.

BACKGROUND OF THE INVENTION

The present invention relates to a technique for dynamically controlling the admission of traffic to a network based in part on the state of the network.

One known type of network is an Internet Protocol ("IP") network. An IP network implements the protocol specified in RFC 791, Internet Protocol <www.cis.ohio-state.edu/htbin/rfc/rfc791.html, visited November 30, 1998>. One type of traffic carried by known IP networks is voice traffic, called Voice over IP ("VoIP") traffic.

FIG. **1** is an example of a prior art VoIP system. It is known to initiate a voice call from a phone set **1** over a conventional circuit-switched network **21** (such as the public switched telephone network (PSTN)) and route the calling party's voice signals to a first gateway **11** connected to the IP network **23**. The first gateway **11** packetizes the voice signals using the Internet Protocol and transmits the packets as VoIP traffic over the IP network **23** to a second gateway **12** closer to the called party than the first gateway **11**. The packets are converted back into voice signals at the second gateway **12**, and those voice signals are routed via the conventional circuit-switched network **22**, to the called party's phone set **2**.

One of the problems with VoIP services is latency. Latency is the delay between the time a signal is sent and the time it is received. Latency adversely affects the quality of service of real-time communications (e.g., voice communications) and is dependent upon the state of the network over which the communications are carried. For example, a heavily burdened network is likely to have more latency than an underutilized network.

A similar problem arises in the context of users making other types of calls over a packet-switched network, such as the Internet. At present, a user can be connected to the Internet by an Internet Service Provider (ISP) and can make a number of calls over the Internet via HTTP (Hypertext Transfer Protocol) commands (using a Web browser such as Microsoft Internet Explorer or Netscape Navigator), FTP (File Transfer Protocol) commands, TELNET connections, and the like. The user may encounter significant delays in accessing, for example, Web sites. Those delays can be caused by a number of factors, including a Web site's inability to respond to all of the users that concurrently seek information from that Web site. A user also may experience significant delays in accessing a particular Web site, not due to that Web site's inability to meet the demand for that site, but due to poor performance characteristics of one or more networks which couple the user to the Web site, or the internetwork routers.

In known VoIP systems, a gateway will pass traffic into a network whenever the gateway has an incoming port that is available to do so. Thus, certain networks must disadvantageously be over-engineered to be able to carry a peak load equal to the traffic that flows when all of the ports of all of the gateways connected to the network are in use. If the traffic sent through the network approaches or exceeds the network's capacity, then the network disadvantageously drops packets (i.e., experiences packet loss) and/or introduces unacceptable delays into communications. In known networks, it is difficult or impossible to guarantee a high quality of service when the network is operating near or at its capacity.

The International Telecommunications Union ("ITU") has established the H.323 standard, which encompasses audio, video and data communications across packet-switched networks, such as the Internet. The H.323 standard was principally developed and established to allow multimedia products and applications from multiple vendors to interoperate. H.323 systems may include a gatekeeper, which can provide bandwidth management. For example, the gatekeeper can reject calls from a terminal if it determines that sufficient bandwidth is not available. H.323 bandwidth management also operates during an active call if a terminal requests additional bandwidth, and the gatekeeper may grant or deny the request for additional bandwidth. Likewise, there are other Internet protocols that provide for establishing or rejecting calls based on bandwidth requirements (e.g., RFC 2211, Specification of the Controlled-Load Network Element Service, <www.cis.ohio-state.edu/htbin/rfc/rfc2211.html, visited Jan. 11, 1999>; RFC 2210, The Use of RSVP with IETF Integrated Services, <www.cis.ohio-state.edu/htbin/rfc/rfc2211.html, visited Jan. 11, 1998>). These bandwidth management protocols do not provide for admitting or rejecting calls based on delay characteristics of the network.

SUMMARY OF THE INVENTION

The present invention provides a system for regulating the call traffic into a packet-switched network based in part upon delay characteristics of the network. In an embodiment of the present invention, a call delay characteristic requirement for a call is determined, a delay characteristic parameter of the packet-switched networks is determined, and a call action based at least partly upon the determined delay characteristic requirement and the determined delay characteristic parameter is performed.

In one embodiment of the invention, the network is an Internet Protocol (IP) network carrying Voice over IP (VOIP) traffic. A voice call made in connection with a VoIP service is not admitted to the IP network and is held if one or more current delay characteristic parameters of the IP network do not satisfy one or more prescribed delay characteristic requirements. Delay characteristic parameters can be periodically updated, and when the current value of one or more delay characteristic parameters satisfy one or more prescribed delay requirements, the VoIP call is admitted to the IP network.

Another embodiment of the present invention dynamically controls the admission of other traffic to an IP network, including multimedia communications, HTTP commands, FTP commands, TELNET connections, and the like. This embodiment allows such data calls to be admitted to the IP network when the IP network satisfies the delay requirements.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG; 1 is an illustration of a prior art Voice over IP system.

FIG. 2 shows a system in accordance with an embodiment of the present invention.

FIG. 3 shows an apparatus in accordance with an embodiment of the present invention.

FIG. 4 shows a system with a quality of service computer in accordance with an embodiment of the present invention.

FIG. 5 is a flowchart illustrating a method in accordance with an embodiment of the present invention whereby a voice call is rerouted over another network when the call quality requirements are not satisfied within a set maximum time.

FIG. 6 is a flowchart illustrating a method in accordance with an embodiment of the present invention that can be utilized as part of a service that markets a certain level of network performance to network users placing calls over the network.

DETAILED DESCRIPTION

A method and apparatus for dynamically controlling the admission of calls to a packet-switched network is described. In the following description, for purposes of explanation, numerous specific details are set forth to provide a thorough understanding of the present invention. It will be obvious, however, to one skilled in the art that the present invention may be practiced without these specific details. In other instances, well known structures and devices are shown in block diagram form and process steps are shown in flowcharts to describe the present invention. Furthermore, it is readily apparent to one skilled in the art that the specific sequences in which steps are presented and performed are illustrative and it is contemplated that the sequences can be varied and still remain within the spirit and scope of the present invention.

A packet-switched network is meant to encompass any network that routes information in the form of discrete packets. For example, a packet may have a header and a payload. The header can include routing and contextual information, which can include the sender's address, the destination address, and a packet number that indicates the packet's place in a series of packets that together comprise a set of information such as a message or a file. Packets can be routed in a connectionless or connection-oriented fashion. In a connectionless protocol such as the Internet Protocol, the path taken by a packet is determined at each router based upon the packet's destination address and conditions in the network. In a connection-oriented protocol, such as the Asynchronous Transfer Mode ("ATM") protocol, a packet is routed along a predetermined path (a circuit) through the network based upon a circuit number that is assigned to the packet.

As used herein, the term "packet-switched network" is meant to encompass connectionless packet-switched networks, connection-oriented networks, and any network that employs a combination of connectionless and connection-oriented protocols to route packets.

As used to describe the present invention, a call is meant to encompass any communication that is carried by a network between entities that are coupled by that network. An entity is anything adapted to utilize a network to communicate with any other entity. Examples of an entity include a telephone, a computer, a facsimile machine, etc. For example, a voice call includes a communication that is carried by a network between a calling party and a called party. As used herein, admitting a call to a network means permitting the network to carry the call.

In an embodiment of the present invention, a calling party initiates a voice call over a conventional circuit-switched network. The incoming voice call is routed over a circuit-switched network from the calling party to a gateway. The gateway decides whether to admit the call to the packet-switched network based in part on the state of the packet-switched network. For example, if the network is already overburdened, the gateway will not admit the call to the network. If, on the other hand, the network can carry the call with an appropriate quality of service, the gateway will admit the call to the network.

The state of a packet-switched network can be indicated by a number of performance parameters, including total delay, mean and standard deviation for such delay, packet loss, error rate, etc. These network characteristic parameters can be determined by methods well known in the art. For example, total delay is the time interval from when one party utters a sound to when the other party hears that sound. It can be determined by methods well known in the art including the timed transmission of audible tones. Packet loss is the percentage of packets transmitted but not received, and can be measured by sending a known set of packets and determining how many are received.

Call quality requirements for the various performance parameters of the packet-switched network (e.g., total delay, mean and standard deviation for such delay, and packet loss) can be established to enable a higher quality of service for certain calls. For example, one delay characteristic requirement is a typical delay requirement, which can require that the network's typical delay parameter be below a certain maximum value before the gateway admits the call to the packet-switched network. As used to describe the call delay characteristic requirements and network delay characteristic parameters, the meaning of the term "typical delay" encompasses an average delay, a mean delay, a median delay, an arithmetic mean delay, a weighted average delay, and other derived delay values that represent a practicable expected delay value. For example, one type of a typical delay requirement is a maximum mean delay requirement, and one type of a typical delay parameter is a mean delay parameter.

Another call delay characteristic requirement is a delay variation requirement, which can require that the network's delay variation parameter be below a certain maximum value before the gateway admits the call to the packet-switched network. As used herein, the meaning of the term "delay variation" encompasses a delay standard deviation, other order moments of the delay distribution, a delay variance, a delay coefficient of skewness, a delay kurtosis, a delay covariance, a delay range, a delay standard error, a delay maximum, a delay minimum, and other derived delay values that represent a practicable delay variation value. For example, one type of a delay variation requirement is a maximum delay standard deviation requirement, and one type of a delay variation parameter is a delay standard deviation parameter.

A call quality requirement (e.g., delay characteristic requirement) may be particular to certain types of calls, call services, the calling party, the called party, and other call differentiations known to one skilled in the art.

Each incoming voice call to the packet-switched network can be held if the current values of the performance parameters are outside the prescribed call quality requirements. As each incoming call is held, actual values of the performance parameters are updated. Various call actions can be taken

while the voice call is held (e.g., sending a wait message to the calling party, sending the calling party a ringing message), and various call actions can be taken if the voice call cannot be admitted to the packet-switched network (e.g., holding the voice call, sending the calling party a busy signal, providing the calling party the option of having the system call him or her back when the VoIP call can be admitted to the network, or rerouting the voice call over another network, such as a conventional circuit-switched network). Voice calls are admitted to the packet-switched network when the current values of the performance parameters are within the prescribed call quality requirements. Once a voice call is admitted to the packet-switched network, all packets associated with the call can be permitted to proceed back and forth through the network as the calling party and called party converse.

While one embodiment of the present invention concerns VoIP services, other embodiments of the present invention concern the admission of any type of call to a packet-switched network. Other types of calls encompassed by the present invention include multimedia communications (e.g., video phone calls), HTTP commands, FTP commands, TELNET connections, and other calls that concern the transmission of data across a packet-switched network. As used to describe the present invention, multimedia communications include audio, video, graphics, animation, facsimile, text communication, and any combination thereof.

FIG. **2** shows a VoIP system which operates in accordance with an embodiment of the present invention. Referring to FIG. **2**, the system includes an admission control gateway **100** that is coupled to a first network **110**, such as a PSTN or a private branch exchange (PBX). Connected to the first network **110** is a telephony station **111**. Examples of such a telephony station include a conventional telephone, a wireless telephone station, a personal computer system with a microphone and headphones, a video conferencing system, a facsimile machine containing a phone handset, etc. The admission control gateway **100** is also coupled to an IP network **140**, which is also connected to a second gateway **180**. A third network **190** is connected to the second gateway **180** and to a telephony station **191**. Networks **110** and **190** may be separate telephone networks or different parts of the same telephone network.

Admission control gateway **100** performs functions that are well known in the art, including receiving from the first network **110** voice signals from a voice call initiated at telephony station **111**, packetizing the voice signals using the Internet Protocol, and transmitting the packets over the IP network **140**. Gateway **180** also performs functions that are well known in the art, including receiving from the IP network **140** packets containing packetized voice signals, converting those packets into voice signals, and routing the voice signals over the third network **190** to the called telephony station **191**. Moreover, gateways **100** and **180** can receive and route other data calls, such as those associated with multimedia communications, HTTP commands, FTP commands, TELNET connections, etc. As used to describe the present invention, the admission control gateway **100** receives one type of a data call when it receives data from the first network **110** to be transmitted over the IP network **140**. A gateway to a packet-switched network also receives a data call when it receives packets of data from another network (e.g., a conventional circuit-switched network, an ATM network, an IP network, etc.) to be transmitted over the packet-switched network. Each gateway can also accumulate data parameters about the network and the current traffic, including network performance parameters, e.g., by polling every other gateway in the network and/or receiving data from network components, such as routers (not shown). Hence, each gateway is able to keep or access up-to-date network data parameters.

The admission control gateway **100** can place dynamic controls on the calls that are to be admitted to the IP network **140** at any given time. Quality of service can be monitored and access controlled to allow calls into the IP network when acceptable service is assured. Call quality for Voice over IP calls can thereby be maintained at an acceptable level. Customer complaints regarding poor quality calls can be reduced. Overengineering of facilities and other resources can be minimized, saving capital and expense.

Referring to FIG. **3**, the admission control gateway **100** includes a processor **250** and a memory **200**. The processor **250** in one embodiment is a general purpose microprocessor, such as the Pentium II processor manufactured by the Intel Corporation of Santa Clara, Calif. In another embodiment, the processor **250** is an Application Specific Integrated Circuit (ASIC), which has been designed to perform in hardware and firmware at least part of the method in accordance with an embodiment of the present invention. Memory **200** is any device adapted to store digital information, such as Random Access Memory (RAM), flash memory, a hard disk, an optical digital storage device, any combination thereof, etc. As shown in FIG. **3**, memory **200** is coupled to processor **250**, a port **282** adapted to be coupled to a sender of a call (e.g., a circuit-switched network), and a port **284** adapted to be coupled to a packet-switched network. The term "coupled" means connected directly or indirectly. Thus, A is "coupled" to C if A is directly connected to C, and A is "coupled" to C if A is connected directly to B, and B is directly connected to C.

In accordance with one embodiment of the present invention, dynamic network call admission instructions are stored on a medium and distributed as software. The medium is any device adapted to store digital information, and corresponds to memory **200**. For example, a medium is a portable magnetic disk, such as a floppy disk; or a Zip disk, manufactured by the Iomega Corporation of Roy, Utah; or a Compact Disk Read Only Memory (CD-ROM) as is known in the art for distributing software. The medium is distributed to a user that has a processor suitable for executing the dynamic network call admission instructions, e.g., to a user with a gateway having a processor, memory, a port adapted to be coupled to a circuit-switched network, and a port adapted to be coupled to a packet-switched network.

Exemplary data structures and instructions adapted to be executed by a processor stored in the memory **200** include the call quality requirements data structure **220**, the packet-switched network performance parameters data structure **230**, and the dynamic call admission instructions **240**.

The call quality requirements data structure **220** can contain the call quality requirements for all calls, certain types of calls, and/or each individual call. For example, a call may have a maximum delay bound, d, for call connection. The delay, d, has mean μ and standard deviation σ. A maximum bound for packet loss, defined by ρ, can also be included in the call quality requirements data structure **220**. Other maximum bounds may be established for error rates and other network performance parameters concerning the IP network, current network traffic and projected network traffic. The call quality requirements data (e.g., the maximum delay bound d) can be predetermined for all calls received by the gateway and stored in the call quality requirements data structure **220**. Alternatively, the call qual-

ity requirements data may be stored in a lookup table that specifies certain call quality requirements for certain types of calls, specific calling parties, specific called parties, etc. For example, the call quality requirements data structure **220** can contain a lookup table indexed according Automatic Number Information (ANI) of the call. The gateway utilizes the calling party's ANI to determine the call quality requirements data for that call from the lookup table. Furthermore, the call quality requirements data can vary for each call by including the call quality requirements data in the call signal itself. The gateway then extracts or reads the call quality requirements data from each call and then stores that data in the call quality requirement data structure **220**.

The packet-switched network performance parameters data structure **230** includes current and/or projected performance parameters of the network. Examples of network performance parameters data include call delay, packet loss, error rate, etc. The delays associated with a call include d_1, d_2, and d_3: d, is the time taken by the incoming gateway to packetize the voice signals; d_2 is the time taken by the outgoing gateway to reassemble the packets into voice signals; and d_3 is the time taken to relay the packets through the IP network and has a standard deviation σ_3. In one embodiment, the incoming and outgoing gateways are the gateways closest to the calling party and the called party, respectively. The delays d_1 and d_2 are functions of $\lambda_i^{(j)}$ and $\lambda_o^{(j)}$, the rate of incoming and outgoing calls at gateway j, respectively. The packet loss p and the delay d_3 are both functions of the IP network and traffic. Although d_1 and d_2 have means μ_1 and μ_2 and standard deviations σ_1 and σ_2 respectively, the standard deviations σ_1 and σ_2 are assumed to be small so that the actual delay, $d_{actual}=(d_1+d_2+d_3)$ has mean $\mu=(\mu_1+\mu_2+\mu_3)$ and standard deviation σ_3.

The network performance parameters data can be accumulated by each gateway such that each gateway keeps up-to-date data. Alternatively, the gateway ascertains the network performance parameters data by accessing quality of service computer **320** of FIG. **4** that determines the appropriate data for each gateway.

Memory **200** stores the dynamic call admission instructions **240** adapted for execution by processor **250**. The term "adapted for execution" is meant to encompass any instructions that are ready for execution in their present form (e.g., machine code) by processor **250**, or require further manipulation (e.g., compilation, decryption, or provided with an access code, etc.) to be ready for execution by processor **250**. The dynamic call admission instructions **240** can determine and indicate when a call has been received by the admission control gateway **100**. The call can then be placed into a queue or otherwise controlled. In some applications of the present invention, the dynamic call admission instructions **240** track how much time has passed since a call was received by the admission control gateway **100** so that the appropriate call action can be taken. In addition, the dynamic call admission instructions **240** can ascertain and make available information about the call such as calling party's ANI, the called party's phone number or IP address, and the type of call requested (e.g., voice call, data call, etc.).

The dynamic call admission instructions **240** also can determine for each call the appropriate call action based upon a call quality requirement and a network performance parameter. For example, a call has maximum bounds for delay and packet loss, which are $d(\mu_{maximum}, \sigma_{maximum})$ and $\rho_{maximum}$, respectively. One call action may be to hold a call in a queue if it is determined that any of the current measured parameters μ, σ_3, and ρ exceed the maximum boundaries. Furthermore, as the call is held in queue, μ, σ_3, and ρ are updated every t_k seconds. The call is admitted to the network when the updated parameters are less than the respective maxima.

In particular, in one embodiment of the present invention, a VoIP call arrives at the gateway at time t_k. The VoIP call is admitted to the IP network based on the following steps:

1. The call characteristic requirements data $\mu_{maximum}$, $\sigma_{maximum}$, and $\rho_{maximum}$, are determined;
2. The network characteristic parameters data $\mu(t_k)$, $\sigma(t_k)$, and $\rho(t_k)$ are determined;
3. If $\mu(t_k)\leqq\mu_{maximum}$, $\sigma(t_k)\leqq\sigma_{maximum}$, and $\rho(t_k)\leqq\rho_{maximum}$, the call is admitted to the IP network;
4. If any one of $\mu(t_k)$, $\sigma(t_k)$, or $\rho(t_k)$ exceeds its respective maximum bound, then the call is held in a queue; and
5. At t_{k+1}, set k+1=k and go to step 1.

In this particular embodiment, the VoIP call is held in the queue until the call is admitted to the IP network. Alternative call actions can provide that the call is held in the queue for a specified amount of time; and if the call has not been admitted to the IP network within that time, then the call is routed over a conventional circuit-switched network to the called party. Another call action can provide for a call back to the calling party when the call can be admitted to the IP network. Instead of holding the call in the queue, the call action can also be to send the calling party a distinctive busy signal or message that indicates that the IP network cannot handle the call at the present moment.

FIG. **4** is an example of another embodiment of the present invention that admits calls from an initiator computer **310** to a contacted computer **390** when the parameters of network **340** satisfy certain requirements. In this embodiment, the calls from the initiator computer **310** can be voice calls and/or data calls such as multimedia communications, HTTP commands, FTP commands, TELNET commands, etc. The admission control gateway **300** receives the call from the initiator computer **310**, determines the call quality requirements, determines the network performance parameters from information provided by a quality of service computer **320**, and takes a call action based on the determined call quality requirements and network performance parameters. The quality of service computer **320** is able to keep up-to-date data parameters about the network, the current traffic, and/or projected traffic by methods well known in the art, including polling every other gateway in the network, receiving data from network components such as routers (not shown), and/or accessing data concerning historical and cyclical traffic patterns (e.g., peak voice call traffic occurs between certain hours weekdays, peak residential data calls occur between certain hours each day, etc.).

FIG. **5** illustrates exemplary steps whereby an embodiment of the present invention reroutes a voice call over another network, such as a conventional circuit-switched network, when the call quality requirements are not satisfied within a set maximum time measured after the voice call is received by the admission control gateway. After an incoming call is received (step **510**), call quality requirements are determined (e.g., a typical delay requirement, a delay variation requirement) (step **520**) and time variable T is set to equal 0 (step **530**). The packet-switched network performance parameters are determined (e.g., a typical delay parameter, a delay variation parameter) (step **540**), and time variable T is incremented (step **550**). The determined network performance parameters are compared to the determined call quality requirements to ascertain whether the call quality requirements are satisfied (step **560**). If the network performance parameters satisfy the call quality

requirements, the call is admitted to the packet-switched network (step **565**). On the other hand, if the call quality requirements are not satisfied, the time variable T is compared to a certain maximum value (step **570**). If T equals or exceeds the maximum value, then the call is routed to a conventional circuit-switched network (step **575**). If T does not equal or exceed the maximum value, then the packet-switched network performance parameters are determined again (step **540**), the time variable T is again incremented (step **550**), etc. Thus, the call is admitted to the packet-switched network if the call quality requirements are satisfied within a certain amount of time, or the call is routed through to a conventional circuit-switched network if the call quality requirements are not satisfied within a certain amount of time.

FIG. **6** illustrates exemplary steps whereby an embodiment of the present invention can be used in conjunction with a service that markets a certain level of network performance to network users placing calls over the network. An account for each user (e.g., a network access user account) may be charged a reduced rate if the network performance for a call is below the marketed level of performance. In particular, and referring to FIG. **6**, after an incoming call is received (step **610**), call quality requirements are determined (e.g., a typical delay requirement, a delay variation requirement) (step **620**); network performance parameters are determined (e.g., a typical delay parameter, a delay variation parameter) (step **630**); and the determined network performance parameters are compared to the determined call quality requirements to ascertain whether the call quality requirements are satisfied (step **640**). If the network performance parameters do not satisfy the call quality requirements, the network access user account is charged a discounted rate (step **655**) and the call is admitted to the network (step **660**). On the other hand, if the call quality requirements are satisfied, the network access user account is charged the standard rate (step **650**) and the call is admitted to the network (step **660**).

In another embodiment of the invention, after a first call action is taken to admit a call to the packet-switched network, a second call action can be performed when the performance parameters of the network no longer satisfy the call quality requirements. For example, periodically during the call (e.g., at specific time intervals) the determined call quality requirements are compared to updated, determined performance parameters to ascertain whether the call quality requirements are still satisfied. When the call quality requirements are no longer satisfied, a second call action can reroute the call over another network. Another second call action may be to charge a discounted rate for the call if the network performance parameters do not satisfy the call quality requirements at a point during the call. Alternatively, the second call action may be to charge a discounted rate for the period of the call during which the network performance parameters do not satisfy the call quality requirements.

Thus, the present invention provides a method and apparatus whereby the admission of calls into a packet-switched network can be dynamically controlled. Performance of the network can be monitored and the admission of a call to the network is controlled to allow the call into the network when a certain level of call quality is met. Call quality can thereby be maintained at an acceptable level.

The invention has been described in conjunction with the preferred embodiment. It is evident that numerous alternatives, modifications, and uses will be apparent to those skilled in the art in light of the foregoing description.

What is claimed is:

1. A method for regulating the admission of a call to a packet-switched network, comprising:

determining a delay characteristic requirement of the call, said requirement including a typical delay requirement and a delay variation requirement;

determining a delay characteristic parameter of the packet-switched network, said parameter including a typical delay parameter and a delay variation parameter; and

performing a call action based at least partly upon whether the determined delay characteristic parameter satisfies the determined delay characteristic requirement, wherein the performing of the call action includes:

admitting the call to the packet-switched network if:

i. the determined typical delay parameter does not satisfy the determined typical delay requirement; or

ii. the determined delay variation parameter does not satisfy the determined delay variation requirement; and

charging a reduced rate for the call.

2. A method for regulating the admission of a call to a packet-switched network, comprising:

determining a delay characteristic requirement of the call, said requirement including a typical delay requirement and a delay variation requirement;

determining a delay characteristic parameter of the packet-switched network, said parameter including a typical delay parameter and a delay variation parameter;

performing a call action based at least partly upon whether the determined delay characteristic parameter satisfies the determined delay characteristic requirement, wherein the performing of the call action includes:

admitting the call to the packet-switched network if:

the determined typical delay parameter satisfies the determined typical delay requirement; and

the determined delay variation parameter satisfies the determined delay variation requirement;

updating the determined typical delay parameter and the determined delay variation parameter after the call is admitted to the network; and

performing a second call action based at least partly upon whether the updated determined typical delay parameter satisfies the determined typical delay requirement and whether the updated determined delay variation parameter satisfies the determined delay variation requirement wherein the performing of a second call action includes charging a reduced rate for the call if:

i. the updated determined typical delay parameter does not satisfy the determined typical delay requirement; or

ii. the updated determined delay variation parameter does not satisfy the determined delay variation requirement.

3. An apparatus for controlling the admission of a call into a packet-switched network, comprising:

a processor; and

a memory, coupled to said processor, storing a plurality of instructions adapted for execution by said processor to:

determine a delay characteristic requirement of the call, said requirement including a typical delay requirement and a delay variation requirement;

determine a delay characteristic parameter of the packet-switched network, said parameter including a typical delay parameter and a delay variation parameter;

perform a call action based at least partly upon whether the delay characteristic parameter satisfies the delay characteristic requirement, wherein said instructions for performing the call action includes instructions to admit the call to the packet-switched network if:

the determined typical delay parameter satisfies the determined typical delay requirement; and

the determined delay variation parameter satisfies the determined delay variation requirement;

update the determined typical delay parameter and the determined delay variation parameter after the call is admitted to the network; and

perform a second call action based at least partly upon whether the updated determined typical delay parameter satisfies the determined typical delay requirement and whether the updated determined delay variation parameter satisfies the determined delay variation requirement wherein said instructions for performing the second call action includes instructions to charge a reduced rate for the call if:

i. the updated determined typical delay parameter does not satisfy the determined typical delay requirement; or

ii. the updated determined delay variation parameter does not satisfy the determined delay variation requirement.

* * * * *

US006222824B1

(12) **United States Patent**
Marin et al.

(10) **Patent No.: US 6,222,824 B1**
(45) **Date of Patent: Apr. 24, 2001**

(54) **STATISTICAL CALL ADMISSION CONTROL**

(75) Inventors: **Gerald A. Marin**, Chapel Hill, NC (US); **Xiaowen Mang**, Metuchen, NJ (US); **Erol Gelenbe**, Durham; **Ralf O. Onvural**, Cary, both of NC (US)

(73) Assignee: **International Business Machines Corporation**, Armonk, NY (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/066,201**

(22) Filed: **Apr. 24, 1998**

(51) **Int. Cl.**[7] **H04J 3/14**; G06F 7/38; H04M 3/00

(52) **U.S. Cl.** **370/230**; 370/252; 370/395; 370/468

(58) **Field of Search** 370/229, 230, 370/235, 236, 238, 252, 254, 389, 395, 468, 477; 709/220, 223, 224, 225, 226; 708/200

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,166,894	*	11/1992	Saito	370/94.1
5,267,232	*	11/1993	Katsube et al.	370/17
5,274,625		12/1993	Derby et al.	370/17
5,289,462		2/1994	Ahmadi et al.	370/60.1
5,335,222	*	8/1994	Kamoi et al.	370/60
5,347,511		9/1994	Gun	370/54
5,359,593		10/1994	Derby et al.	370/17
5,434,848		7/1995	Chimento, Jr. et al.	370/17
5,850,385	*	12/1998	Esaki	370/216
5,914,936	*	6/1999	Hatono et al.	370/230
6,041,039	*	3/2000	Kilkki et al.	370/230
6,067,287	*	5/2000	Chung-Ju et al.	370/232
6,134,239	*	10/2000	Heinanen et al.	370/412

OTHER PUBLICATIONS

Gelenbe, et al., Diffusion based statistical call admission control in ATM, *Performance Evaluation*, vol. 27 & 28, pp. 411–436 (Elsevier Science B.V., 1996).

Rege, Kiran, Equivalent Bandwidth and Related Admission Criteria for ATM Systems—A Performance Study, *International Journal of Communication Systems*, vol. 7, pp. 181–197 (John Wiley & Sons, 1994).

Guerin, Roch, Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks, *IEEE Journal on Selected Areasin Communications*, vol. 9, No. 4, pp. 968–981 (Sep. 1991).

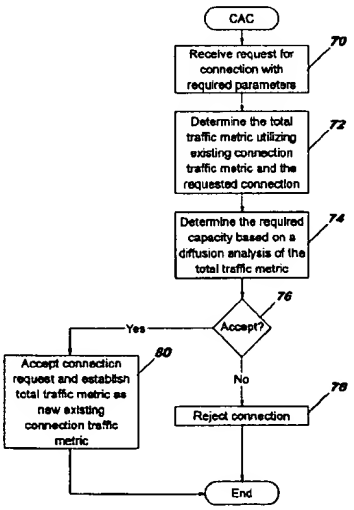
(List continued on next page.)

Primary Examiner—Alpus H. Hsu
(74) *Attorney, Agent, or Firm*—Timothy J. Sullivan; Myers, Bigel, Sibley & Sajovec

(57) **ABSTRACT**

Methods, systems and computer program products are provided for evaluating requests for a network connection combining the requested network connection with existing connections so as to provide a representation of the total network connections and accepting the request for a network connection if sufficient resources are available to accept the request for a network connection based upon the representation of the total network connections. Particular embodiments of the present invention determine if sufficient resources are available to accept the request for a network connection based upon diffusion based representations of the existing network connections and the requested network connection and accepting the request for a network connection if sufficient resources are available to accept the request for a network connection. The diffusion based representation may be used to determine the capacity required for existing connections and the requested connection based upon a predefined maximum loss ratio (L).

42 Claims, 3 Drawing Sheets



OTHER PUBLICATIONS

Feldmeier, David, A Framework of Architectural Concepts for High-Speed Communication Systems, *IEEE Journal on Selected Areas in Communications*, vol. 11, No. 4, pp. 480–488 (May 1993).

Guerin, et al., A Unified Approach to Bandwidth Allocation and Access Control in Fast Packet-Switched Networks, *INFOCOM '92*, pp. 1A.1.1–1A.1.12 (1992).

Gelenbe, Erol, Probabilistic Models of Computer Systems, Part II: Diffusion Approximations, Waiting Times and Batch Arrivals, *Acta Informatica*, vol. 12, pp. 285–303 (1979).

Gelenbe, et al., The Behavior of a Single Queue in a General Network, *Acta Informatica*vol. 7, pp. 123–126 (1976).

Gelenbe, Erol, On Approximate Computer System Models, *Journal of the Association for Computing Machinery*, vol. 22, No. 2, pp. 261–269 (Apr. 1975).

Kobayashi, H., Application of the Diffusion Approximation to Queuing Networks, I: Equilibrium Queue Distributions, *Journal of the Association for Computer Machinery*, vol. 21, No. 2. [[/ 316=328 (Apr. 1974).

Kobayashi, H., Application of the Diffusion Approximation to Queuing Networks II: Nonequilibrium Distributions and Applications to Computer Modeling, *Journal of the Associationi for Computing Machinery*, vol. 21, No. 3, pp. 459–469 (Jul. 1974).

* cited by examiner

FIG. 1

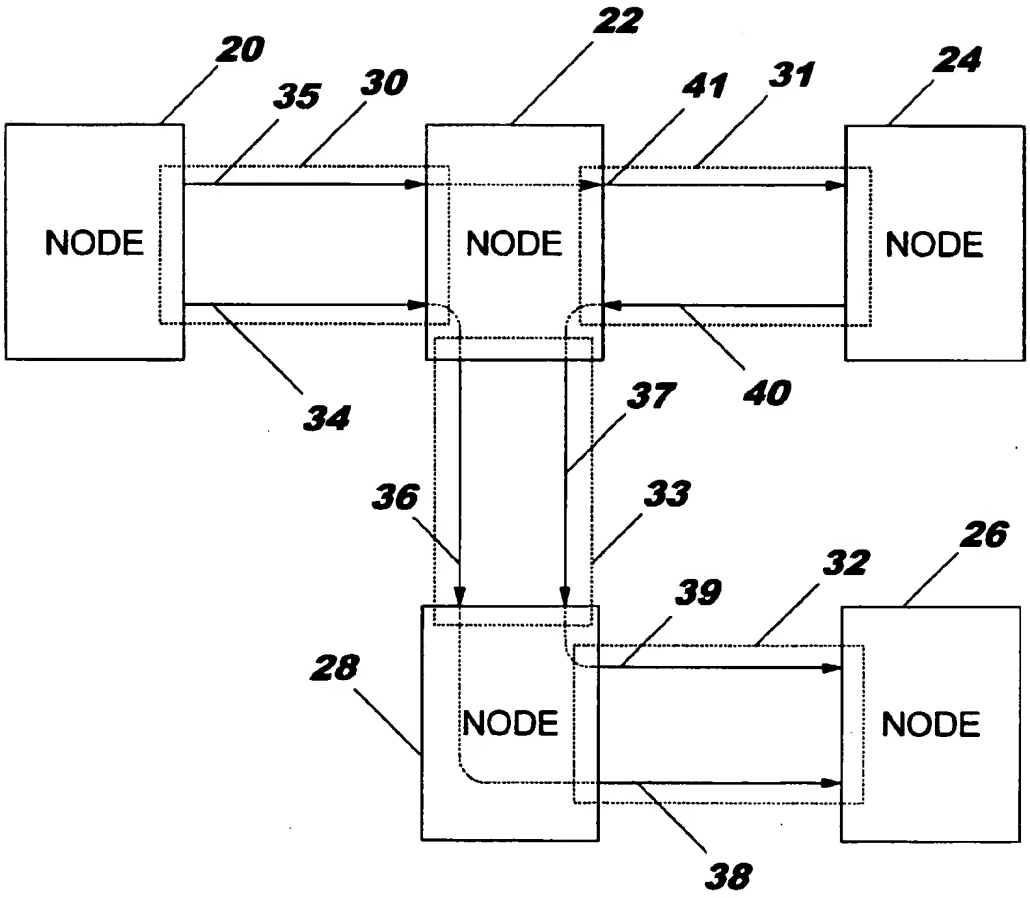

20
22
24
35
30
41
31
NODE
NODE
NODE
34
40
37
36
33
26
32
39
28
NODE
NODE
38

FIG. 2

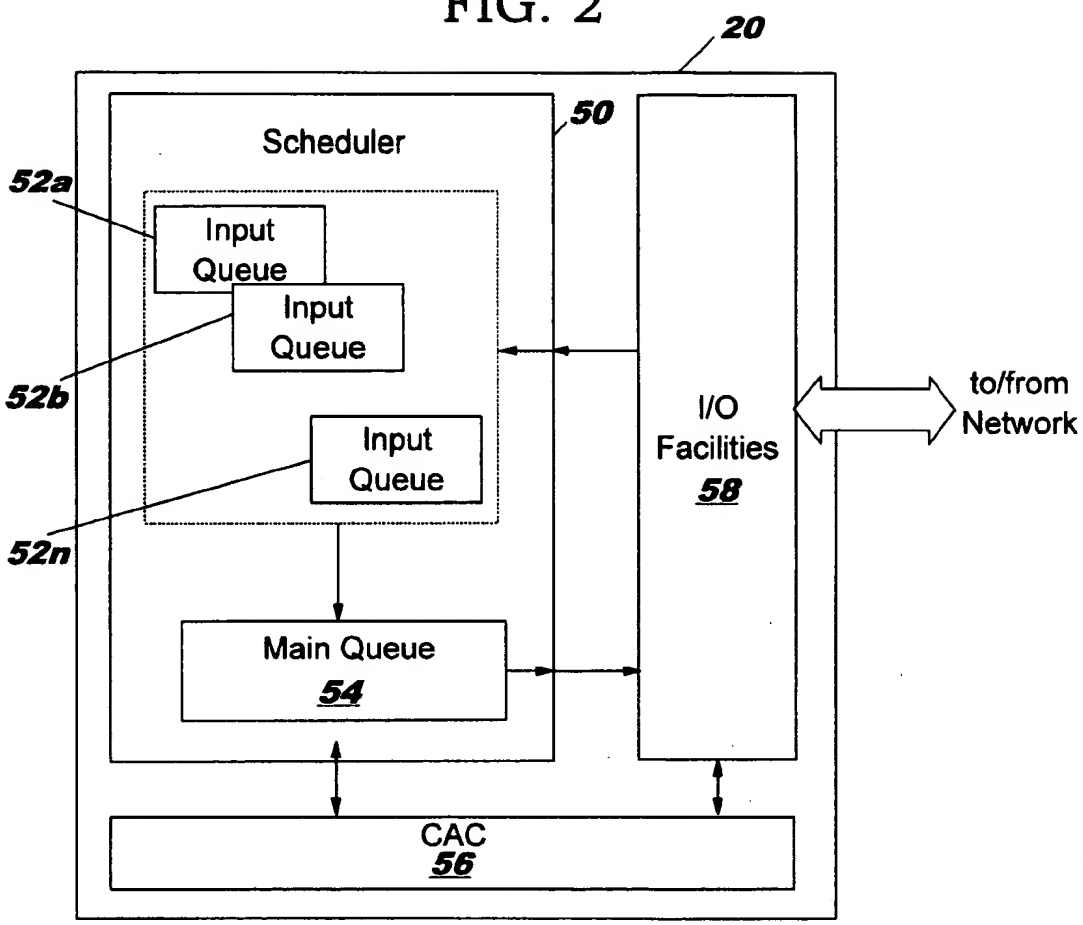
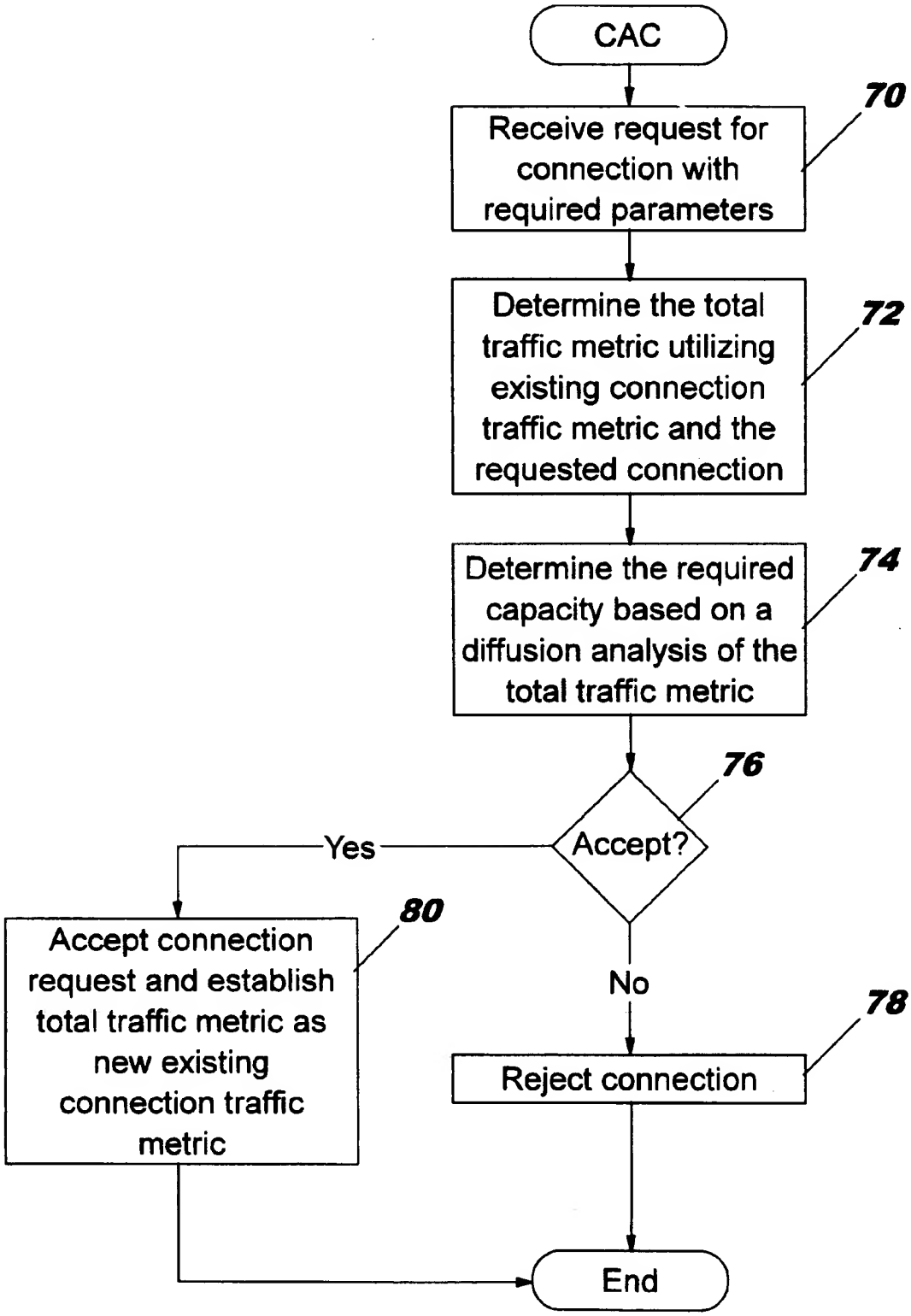

20
50
Scheduler
52a
Input Queue
Input Queue
52b
Input Queue
52n
Main Queue
54
I/O Facilities
58
to/from Network
CAC
56

FIG. 3


CAC
Receive request for connection with required parameters
70
Determine the total traffic metric utilizing existing connection traffic metric and the requested connection
72
Determine the required capacity based on a diffusion analysis of the total traffic metric
74
Accept?
76
Yes
No
Accept connection request and establish total traffic metric as new existing connection traffic metric
80
Reject connection
78
End

STATISTICAL CALL ADMISSION CONTROL

RELATED APPLICATIONS

This application is related to and claims priority from U.S. Provisional Patent Application Ser. No. 60/044,105 filed Apr. 24, 1997 and entitled STATISTICAL BANDWIDTH ALLOCATION IN ATM NETWORKS.

FIELD OF THE INVENTION

The present invention relates generally to communications networks and more particularly to communications networks using Asynchronous Transfer Mode (ATM).

BACKGROUND OF THE INVENTION

Asynchronous Transfer Mode (ATM) networks have become increasingly popular for both wide area networks and local area networks. In an ATM network all information is transported through the network in relatively short blocks called cells. Information flow through the network is along paths called virtual channels which are set up using a series of tables implemented in switching nodes that comprise the network. Cells on a particular virtual channel follow the same path through the network and are delivered to the destination in the order in which they are received.

When communications are initiated in an ATM network a request is made for a connection. As part of the connection request, the quality of service (QoS) of the request is specified by the requestor. Quality of service characteristics which may be specified in ATM include cell transfer delay (network latency), cell delay variation (jitter), cell transfer capacity (average and peak allowed rates), cell error ratio, cell loss ratio and cell misinsertion ratio. These QoS parameters may be used by the ATM nodes to determine the route of the connection and in determining allowable node utilization.

Routing in an ATM network is performed by an ATM node which attempts to find a feasible route for a virtual connection from a source to a destination. An ATM connection is not set up unless and until the network determines that there is sufficient capacity for the new connection. This determination is based upon the characteristics of the network, the existing traffic on the network and the requested QoS for the connection. If the requested QoS cannot be provided then the requested connection is not accepted. The function in ATM which determines whether a connection request is accepted is referred to as call admission control (CAC).

In ATM, the CAC function is typically carried out at two places; the entry point into the ATM network and at each node in the connection path. At the entry point into the ATM network the CAC function determines a feasible route for the connection request based on the QoS requested, and either accepts or rejects the request based on this determination. To perform the entry point CAC function, the entry point should have available information about the current utilization of the other nodes in the ATM network. This information may then be utilized to determine if the nodes along the selected route can satisfy the QoS requirements of the requested connection.

At each node in a connection route, a local CAC function determines if the node will accept the connection request. This CAC function utilizes the requested QoS and information about the existing connections through the node to determine if the requested QoS level may be achieved. If so, the connection request is accepted.

As is apparent from the above discussion, the call acceptance procedure of an ATM system may impact on the performance of the network. If too many connection requests are accepted then the QoS of the connections may be degraded and additional, possibly costly, resources may be required to handle the network's traffic. If too few connection requests are accepted, then the network will be under-utilized. The call admission problem is further complicated by the fact that the call admission procedure is typically carried out in real-time as call requests are received. The limits placed on the complexity of the call admission procedure by the real-time requirement may potentially result in compromises resulting in either too conservative a call admission procedure or too optimistic a procedure. Finally, the call admission problem is also limited by the information available to determine the characteristics of traffic through the network.

If the network makes a decision to admit a new connection, resources such as link bandwidth and buffers must be reserved to guarantee that the connection receives its guaranteed QoS. For certain types of traffic with well-known characteristics (such as voice traffic) it is straightforward to determine the amount of bandwidth needed to guarantee QoS. Where traffic is more variable (bursty), however, the network must make a more difficult decision regarding the appropriate level of bandwidth resources to set aside for the new connection. The equivalent bandwidth of a connection is defined as an "average" (or steady-state) amount of bandwidth needed to be reserved to carry the traffic of that connection when it is sharing link resources with other connections that have been similarly accommodated. CAC mechanisms based on equivalent bandwidth are typically simple in that the determination of whether a given set of connections can be accommodated without violating their QoS requirements reduces to comparing the sum of the equivalent bandwidths of individual sources to the link capacity.

Although generally simple to implement, equivalent bandwidth CAC functions are highly conservative when the buffer size is small or moderate. Thus, utilization of an equivalent bandwidth approach to call admission may result in fewer connection requests being accepted than could be accommodated by an ATM network or link in an ATM network. Accordingly, additional resources in the network may be required to handle the network traffic than would otherwise be needed if the network resources were more efficiently utilized.

A second approach to call admission control based on bandwidth is the Gaussian approximation based on a zero-buffer assumption. If the number of sources being multiplexed (N) is sufficiently large, the aggregate traffic can be approximated by a Gaussian process with mean rate

$$\lambda = \sum_{u=1}^{N} \lambda_u$$

and variance

$$\sigma^2 = \sum_{u=1}^{N} \sigma_u^2.$$

While call admission based on a Gaussian process may provide increased efficiency over an equivalent bandwidth approach, in a system having a buffer, the buffer's capacity

to absorb traffic bursts is ignored, thus, resulting in under-utilization of the network. Furthermore, when N is small, the Gaussian approximation will not be valid which may result in extremely conservative bandwidth determinations.

Various other hybrid systems have also been proposed, however, these systems may also have limitations. For example, a system utilizing a highly non-linear function of the individual equivalent bandwidths to determine the admissibility of a given set of traffic sources may be over-optimistic in certain situations, thus, resulting in higher cell loss ratios than the specified QoS of the connections.

In light of the above discussion, a need exists for improvements in the mechanisms for accepting connections in ATM and other networks.

SUMMARY OF THE INVENTION

In view of the above discussion, it is an object of the present invention to provide for call acceptance which can efficiently utilize network resources.

A further object of the present invention is to provide a call admission procedure which may be carried out in real time.

Still another object of the present invention is to provide a call admission procedure which is not overly optimistic in accepting connection requests.

These and other objects of the present invention are provided by methods, systems and computer program products for evaluating requests for a network connection combining the requested network connection with existing connections so as to provide a representation of the total network connections and accepting the request for a network connection if sufficient resources are available. Particular embodiments of the present invention determine if sufficient resources are available to accept the request for a network connection based upon diffusion based representations of the existing network connections and the requested network connection. The diffusion based representation may be used to determine the capacity required for existing connections and the requested connection based upon a predefined maximum loss ratio (L). In particular embodiments of the present invention, a diffusion based representation of the network resources required by the existing network connections and the requested network connection is generated.

By utilizing a diffusion based representation of network traffic, increased efficiency in the acceptance procedure may be achieved. The use of the diffusion approximation may more accurately reflect actual bandwidth usage of connections thus allowing for more accurate determinations of the impact of an additional connection on the total traffic. Use of the diffusion approximation conservatively estimates required bandwidth so as to assure that the service requirements of existing connections are maintained. However, the diffusion approximation can also allow more connections to be established than would be established using conventional techniques. Furthermore, the diffusion based representation is not so complex as to preclude real-time implementation.

In one embodiment of the present invention, a diffusion based representation of the network resources required by the existing network connections and the requested network connection is achieved by summing an arrival rate of each existing connection and the arrival rate of the requested network connection to provide a total arrival rate (λ), summing a variance of arrival rate of each existing connection and the variance of arrival rate of the requested network connection to provide a total variance (σ^2), and summing an instantaneous variance of the change of the buffer occupancy for each existing connection so as to provide a total instantaneous variance (α). The link capacity required by the existing network connections and the requested network connection is then determined based on the total arrival rate, the total variance and the total instantaneous variance.

In another embodiment of the present invention, the total instantaneous variance may be determined by summing a squared coefficient of variation of the incoming traffic of each existing connection and the squared coefficient of variation of the requested network connection to provide a total squared coefficient of variation (c^2). In such a case determination of the link capacity required by the existing network connections and the requested network connection may be based on the total arrival rate, the total variance and the total squared coefficient of variation.

In particular embodiments, where the network is characterized by a maximum bandwidth (B) and a cell loss ratio (L) the capacity (C*) required by the existing connections and the requested connection may be determined by evaluating:

$$C^*=\lambda-\beta+\sqrt{\beta^2-2\sigma^2\omega_1}$$

where

$$\beta=\frac{2B}{\alpha}\sigma^2$$

and where $\omega_1=\ln(L\sqrt{2\pi})$. Alternatively, the required capacity C* may be determined by by evaluating:

$$C^*=\lambda-\beta+\sqrt{\beta^2-2\sigma^2\omega_2}$$

where

$$\beta=\frac{2B}{\alpha}\sigma^2$$

and where $\omega_2=\ln(L\lambda\sqrt{2\pi})-\ln(\sigma)$. In either case, α may be specified as $\alpha=\lambda c^2$.

Preferably, the network comprises an ATM network having a plurality of ATM nodes. In an ATM network, the present invention may be utilized for all CAC functions. Thus, the present invention may make an acceptance determination for a single node or for each node in the ATM network in a path through the network associated with the requested network connection.

As will be appreciated by those of skill in the art, the present invention may be embodied as methods, apparatus or computer program products.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram schematically illustrating an ATM network utilizing the present invention;

FIG. 2 is a block diagram of an ATM device according to the present invention; and

FIG. 3 is a flowchart illustrating the operation of the present invention in a CAC function.

DETAILED DESCRIPTION OF THE INVENTION

The present invention now will be described more fully hereinafter with reference to the accompanying drawings, in which preferred embodiments of the invention are shown. This invention may, however, be embodied in many different forms and should not be construed as limited to the embodi-

ments set forth herein; rather, these embodiments are provided so that this disclosure will be thorough and complete, and will fully convey the scope of the invention to those skilled in the art. Like numbers refer to like elements throughout.

As will be appreciated by one of skill in the art, the present invention may be embodied as a method, data processing system, or computer program product. Accordingly, the present invention may take the form of an entirely hardware embodiment, an entirely software embodiment or an embodiment combining software and hardware aspects. Furthermore, the present invention may take the form of a computer program product on a computer-readable storage medium having computer-readable program code means embodied in the medium. Any suitable computer readable medium may be utilized including hard disks, CD-ROMs, optical storage devices, or magnetic storage devices.

Operations for various aspects of the present invention are illustrated herein in flowchart illustrations. It will be understood that each block of the flowchart illustrations, and combinations of blocks in the flowchart illustrations, can be implemented by computer program instructions. These computer program instructions may be provided to a processor or other programmable data processing apparatus to produce a machine, such that the instructions which execute on the processor or other programmable data processing apparatus create means for implementing the functions specified in the flowchart block or blocks. These computer program instructions may also be stored in a computer-readable memory that can direct a processor or other programmable data processing apparatus to function in a particular manner, such that the instructions stored in the computer-readable memory produce an article of manufacture including instruction means which implement the functions specified in the flowchart block or blocks.

The computer program instructions may also be executed by a processor or other programmable data processing apparatus to cause a series of operational steps to be performed by the processor or other programmable apparatus to produce a computer implemented process such that the instructions which execute on the computer or other programmable apparatus provide steps for implementing the functions specified in the flowchart block or blocks.

Accordingly, blocks of the flowchart illustrations support combinations of means for performing the specified functions, combinations of steps for performing the specified functions and program instruction means for performing the specified functions. It will also be understood that each block of the flowchart illustrations, and combinations of blocks in the flowchart illustrations, can be implemented by special purpose hardware-based computer systems which perform the specified functions or steps, or by combinations of special purpose hardware and computer instructions.

The present invention utilizes a diffusion based statistical call admission procedure. Preferably, the present invention is utilized in an ATM multiplexer where cells from different connections are interleaved at a transmission buffer and served in a first-come-first-serve manner. Details of such as system, while not essential to an understanding of the present invention, may be found in U.S. patent application Ser. No. 08/968,201 entitled SYSTEMS, METHODS AND COMPUTER PROGRAM PRODUCTS FOR CONTROLLING EARLIEST DEADLINE FIRST SCHEDULING AT ATM NODES.

Utilizing the diffusion based statistical analysis of traffic, whether an acceptable cell loss ratio will result if an additional connection is added to the traffic may be determined. Thus, based on a predefined acceptable cell loss ratio, a determination may be made as to whether to accept a new connection or to reject the connection.

As an example of a diffusion based representation, ATM traffic may be characterized by the following three equations:

$$f(x,t)dx:Pr\{x \leq X(t) < x+dx\},$$

$$m(t):Pr\{X(t)=0\},$$

and

$$M(t):Pr\{X(t)=B\}$$

where X(t) is a random variable denoting buffer size at time t, and B is the total buffer size. In the steady state (as t goes to infinity), dropping the dependency on t results in the following:

$$-\frac{\partial}{\partial x}f(x)+\frac{\alpha}{2}\frac{\partial^2}{\partial x^2}f(x)+\frac{m}{E[h]}\delta(x-1)+\frac{M}{E[H]}\delta(x-B+1)=0 \quad (1)$$

$$\lim_{x\to 0^+}\left[-\mu f(x)+\frac{\alpha}{2}\frac{\partial}{\partial x}f(x)\right]=\frac{m}{E[h]} \quad (2)$$

$$\lim_{x\to B^+}\left[-\mu f(x)+\frac{\alpha}{2}\frac{\partial}{\partial x}f(x)\right]=\frac{M}{E[H]} \quad (3)$$

$$m+M+\int_{0^+}^{B^-}f(x)dx=1$$

where δ(x) is the Dirac Delta function, h is a random variable denoting the distribution of idle period in the queue with expected value E[h], H is a random variable denoting the distribution of time the buffer is full with expected value E[H], and μ is the instantaneous average rate of change of the buffer occupancy which may be determined by the mean aggregate cell arrival rate to the buffer (λ) minus the transmission capacity of the lin (C), ie. $\mu=\lambda-C$.

Equation (1) represents the stationary behavior for the motion of the queue length process in the interval]0,B[and the effects ofjumps from 0 and B into the interval. Equation (2) corresponds to the depletion of the probability mass m at the lower boundary (i.e. when the queue is empty) due to jumps to having one cell in the queu and the flow of probability mass from inside the interval]0,B[towards the lower boundary. Similarly, equation (3) represents the depletion of the probability mass M at the higher boundary (i.e. when the queue is full) due to jumps to having B−1 cells in the queue and the flow of probability mass from inside the interval]0,B[towards the higher boundary.

Using equations (1), (2) and (3), two equations may be derived to approximate cell loss ratio in an ATM'multiplexer. These two equations are a finite buffer approximation (L_{FB}) and an infinite buffer approx (L_{IB}). The equations are as follows:

$$L_{FB} = \psi e^{\frac{2(B-1)}{\alpha}\mu} Pr\{R(t) \geq C\} \quad (4)$$

$$L_{IB} = \gamma e^{\frac{2B}{\alpha}\mu} \frac{E[(R(t) \geq C)^+]}{\lambda} \quad (5)$$

where

$$\psi = \frac{-\mu E[H]}{(1-\mu E[h]) - (1+\mu E[H])e^{\frac{2(B-1)}{\alpha}\mu}}$$

and where

$$\gamma = \frac{1}{1-\mu E[h]}\left[1 - e^{\frac{2\mu}{\alpha}}\right]\frac{\alpha}{2\mu} \text{ with } R(t)$$

denoting the instantaneous cell arrival rate at time t.

These approximations may be used in the call admission control procedure to determine whether or not a new connection request can be accepted if there are N connections already established and N is great than or equal to 0. Either equation (4) or equation (5) may be used to provide the cell loss ratio when there are N connections multiplexed on a link with transmission capacity C. The amount of bandwidth required to support a new connection request when there are N connections already established at the link is equivalent to determining he minimum value of C so that the right hand side of equation (4) or equation (5) is lee than or equal to the desired cell loss ratio.

Considering that call admission decisions are required to be performed in real time and that equations (4) and (5) do not yield a closed form solution for C, a conservative approximation to estimate the value of C may De derived as follows for the finite buffer case (CFB):

$$C_{FB}=\lambda-\beta+\sqrt{\beta^2-2\sigma^2\omega_1} \quad (6)$$

where

$$\beta = \frac{2B}{\alpha}\sigma^2$$

and where $\omega_1=\ln(L\sqrt{2\pi})$ for a desired cell loss ratio of L.

Similarly, for the infinite buffer case a conservative approximation to estimate the value of C (C_{IB}) may be derived as follows:

$$C_{IB}=\lambda-\beta+\sqrt{\beta^2-2\sigma^2\omega_2} \quad (7)$$

where $\omega_2=\ln(L\lambda\sqrt{2\pi})-\ln(\sigma)$.

A new connection is established if there is enough bandwidth to accommodate the connection. Either equation (6) or equation (7) may be determined for the total connections (i.e. the existing connections and the new connection) and if the required capacity is less than the capacity of the link, then the connection request is accepted.

The above description generally describes the development of a diffusion based representation of the network traffic which may be used in the call admission control procedure. As used herein, diffusion based representation is used to refer to representation of the network traffic based on a diffusion approximation of the traffic. A diffusion approximation, typically, approximates a discrete process with a continuous process based on the central limit theorem. The diffusion approximation results in the new connection request being combined with existing requests so as to provide a representation of the network connections which incorporates the requested connection and the existing connections.

A specific implementation of the present invention will now be described with reference to FIGS. 1 through 3. FIG. 1 illustrates a network utilizing ATM nodes according to the present invention. As seen in FIG. 1, a number of ATM nodes **20**, **22**, **24**, **26** and **28**, are interconnected via physical links **30**, **31**, **32** and **33**. The physical links provide for communication between the nodes and allow for the interconnection of source and destination ATM nodes via an end-to-end virtual circuit formed by interconnected, but separate inter-node segments. FIG. **1** are illustrated a number of such segments **34**, **35**, **36**, **37**, **38**, **39**, **40** and **41**.

As an example of a path between nodes **20** and **26**, a virtual circuit between the two nodes may be "constructed" from logical connection **34** over physical link **30**. logical connection **36** over physical link **33**, and logical connection **38** over physical link **32** to node **26**.

As will be appreciated by those of skill in the art, the nodes of FIG. 1 may be ATM endpoint nodes, ATM switches, user network interface nodes or other processing systems utilizing the ATM communication protocols. Thus, the present invention is not limited to use in ATM switches but may be utilized in any devices complying with ATM standards. Furthermore, while the network of FIG. **1** has been described as an ATM network, as will be appreciated by those of skill in the art, the network may be any packet based network utilizing a call admission procedure where sufficient information is available to utilize the diffusion based statistical call admission procedure of the present invention.

When a request by a user to access the ATM network of FIG. 1 is made, for example, to node **20**, the node evaluates whether the quality of service parameters of the request may be met by a route through the ATM network before the connection request is accepted. Thus, for example, if a request is made for a connection between node **20** and node **24**, node **20** would evaluate the status of connections at node **20**, node **22**, and node **24** to determine whether to accept the request. Thus, node **20** will typically have information regarding the connections which already exist through node **22** and node **24**. The CAC function of node **20** will then determine whether the request may be accepted utilizing the diffusion based call admission control of the present invention to evaluate whether each node in the connection path may accept the connection request.

FIG. **2** illustrates one embodiment of an ATM node according to the present invention. As seen in FIG. **2**, an ATM node according to the present invention may include a scheduler **50**, input queues **52***a*, **52***b* through **52***n*, a main queue **54**, a CAC function **56** and input/output capabilities **58**. Cells are received and transmitted through the node by the input/output facilities **58**. Received cells are placed in the appropriate input queue **52***a* through **52***n* by the scheduler **50**. Cells that are to be output are moved by the scheduler **50** from the input queues to the main queue **54** for transmission. The CAC function **56** according to the present invention determines whether requests for connections to the node will be accepted (the ACAC function) utilizing the diffusion based call admission control procedure described herein. The CAC **56** also determines a feasible route for connection requests received by the node when the node serves as an entry point to the ATM network.

The present invention is described with respect to the nodes and networks of FIG. 1 and FIG. 2; however, as will be appreciated by those of skill in the art, the present invention is applicable to any network configuration where call acceptance is performed. Furthermore, the functions of the present invention may be implemented utilizing a data processing system operating under software control, as dedicated hardware or a combination of the two.

FIG. 3 illustrates one embodiment of the present invention. As is seen in that figure, the CAC function begins when a request is received by an ATM node for a connection (block 70). For example, a request for a connection may be received at node 20 in FIG. 1 which may be an ATM device such as is illustrated in FIG. 2. The request will include parameters which define the characteristics of the requested connection. For example, the request may contain information regarding packet arrival rates (i.e. mean aggregate cell arrival rate λ), variance in packet arrival rate (σ^2) and the instantaneous variance of the change of the buffer occupancy (α) or such information from which this information could be derived. For example, the instantaneous variance of the change of the buffer occupancy may be determined by $\alpha=\lambda^3V_a+C^3V_s$ where V_a is the variance of the interarrival time, V_s is the variance of service time and C is the transmission capacity of the link. Alternatively, for a specific connection, the instantaneous variance of the change of the buffer occupancy for the connection may be based on a model of the connection. For example, for an "on-off" connection, $\alpha=\lambda c^2$ may be utilized where c^2 is the squared coefficient of variation of the incoming traffic from the connection.

After the connection request is received, a total trffic metric is determined based on traffic parameters of the existing connections and the requested connection (block 72). The traffic metric may be determined by adding the traffic metric for the requested connection to the traffic metric for the existing connections. Thus, for example. an existing traffic metric I for the existing connections may be expressed as

$$I=\left\{\sum_{u=1}^{N}\lambda_u,\ \sum_{u=1}^{N}\sigma_u^2,\ \sum_{u=1}^{N}\alpha_u\right\}.$$

Similarly, the traffic metric U for the requested connection may be expressed as

$$U=\{\lambda_u,\ \sigma_u^2,\ \alpha_u\}.$$

Thus the total traffic metric T may be expressed as T=I+U. As will be appreciated by those of skill in the art in light of the above discussion, the information to produce the total traffic metric may be based on traffic models or on measurements of traffic parameters or a combination of the two wherein for certain connections the parameters are derived from models and for other connections the parameters are empirically derived from traffic measurements. If, for example, the parameters are based on a traffic model, then a may be expressed as $\alpha=\lambda c^2$ which would result in I being defined as

$$I=\left\{\sum_{u=1}^{N}\lambda_u,\ \sum_{u=1}^{N}\sigma_u^2,\ \sum_{u=1}^{N}\lambda_u c_u^2\right\}.$$

U may then be defined as

$$U=\{\lambda_u,\ \sigma_u^2,\ c_u^2\}$$

from which the total traffic metric T may be determined. While the present invention has been described with reference to two methods of determining parameters of the traffic metric for a connection, as will be appreciated by one of skill in the art, various other ways may be utilized to determine values of the parameters of the traffic metric for a connection or requested connection.

After the total traffic metric is determined, this information is then used to calculated a required capacity (C*) based on the parameters of T. C* may be determined utilizing equation (6), equation (7) or both. Thus, in block 74 the ATM node determines one or both of the following:

$$C_{FB}^{*}=\lambda-\beta+\sqrt{\beta^2-2\alpha\omega_1}\ \text{ or }\ C_{IB}^{*}32\ \lambda-\beta+\sqrt{\beta^2-2\alpha\omega_2}$$

where

$$\beta=\frac{2B}{\alpha}\sigma^2,$$

where $\omega_1=\ln(L\sqrt{2\pi})$ and where $\omega_2=\ln(L\lambda\sqrt{2\pi})-\ln(\pi)$ for the cell loss ratio (L) specified for the network and utilizing the parameters from the total traffic metric T described above. As is described above, α may vary depending on the particular implementation and may be $\alpha=\lambda^3V_a+C^3V_s$ or may be $\alpha=\lambda c^2$.

Once C* is determined for the total traffic metric, it is determined if the requested connection will be accepted (block 76). This determination may be made by comparing C* with the capacity of the link (C). If C* is greater than C then the connection request is not accepted and if C* is less than or equal to C then the connection is accepted. Alternatively, in a more conservative approach, if C* is greater than or equal to C then the connection request is not accepted and if C* is less than C then the connection is accepted. If both equations are utilized then the comparison of C* with C for both results may be made in which case both of the results may be used in determining if the connection request should be accepted.

If the connection is accepted, then conventional procedures are utilized to accept the connection and the existing traffic metric I is set to the total traffic metric T to reflect the new connection (block 80). If the connection is rejected then the conventional procedure for rejection of a connection is carried out (block 78). In such a case, the existing traffic metric I remains unchanged.

While the present invention has been described with respect to a particular series of operations in a CAC function, as will be appreciated by those of skill in the art other series of operations could be utilized while still benefiting from the teachings of the present invention. Furthermore, as will be appreciated by those of skill in the art, the present invention may be applicable to networks other than ATM networks so as to achieve the benefits and advantages of the present invention in networks other than ATM networks. The invention may advantageously be used in any network that sets aside resources to support QoS for connections in part through implementation of admission control processes.

The foregoing is illustrative of the present invention and is not to be construed as limiting thereof. Although a few exemplary embodiments of this invention have been

described, those skilled in the art will readily appreciate that many modifications are possible in the exemplary embodiments without materially departing from the novel teachings and advantages of this invention. Accordingly, all such modifications are intended to be included within the scope of this invention as defined in the claims. In the claims, means-plus-function clause are intended to cover the structures described herein as performing the recited function and not only structural equivalents but also equivalent structures. Therefore, it is to be understood that the foregoing is illustrative of the present invention and is not to be construed as limited to the specific embodiments disclosed, and that modifications to the disclosed embodiments, as well as other embodiments, are intended to be included within the scope of the appended claims. The invention is defined by the following claims, with equivalents of the claims to be included therein.

That which is claimed is:

1. A method of evaluating requests for a network connection, the method comprising the steps of:

determining if an acceptable cell loss ratio results from a requested network connection being added to network traffic utilizing a diffusion based statistical analysis of the network traffic; and

accepting the request for a network connection if an acceptable cell loss ratio results from a requested network connection being added to the network traffic based on the diffusion based statistical analysis.

2. A method according to claim 1, wherein said step of determining comprises the step of generating a diffusion based representation of the network traffic for a predefined maximum acceptable cell loss ratio resulting from the existing network connections and the requested network connection; and

wherein said step of accepting comprises the step of accepting the request for network access if the diffusion based representation of the network traffic for the existing network connections and the requested network connection is less than the capacity of the network.

3. A method according to claim 2, wherein said generating step generates a diffusion based representation of the capacity required for existing connections and the requested connections.

4. A method of evaluating requests for a network connection, the method comprising the steps of:

determining if sufficient network capacity is available to accept the request for a network connection based upon diffusion based representations of network capacity for the existing network connections and the requested network connection; and

accepting the request for a network connection if sufficient network capacity is available to accept the request for a network connection based upon the diffusion based representation of the existing network connections and the requested network connection.

5. A method according to claim 4, wherein said determining step determines the capacity required for existing connections and the requested connections based upon a cell loss ratio (L).

6. A method according to claim 4, wherein said step of determining comprises the step of generating a diffusion based representation of the network capacity required by the existing network connections and the requested network connection; and

wherein said step of accepting comprises the step of accepting the request for network access if the diffusion based representation of the network capacity required by the existing network connections and the requested network connection is less than the capacity of the network.

7. A method of determining whether to accept a request for a network connection, the method comprising the steps of:

summing an arrival rate of each existing connection and the arrival rate of a requested network connection corresponding to the request for a network connection to provide a total arrival rate (λ);

summing a variance of arrival rate of each existing connection and the variance of arrival rate of the requested network connection to provide a total variance (σ^2);

summing an instantaneous variance of a change of the buffer occupancy for each existing connection so as to provide a total instantaneous variance (α);

determining a link capacity required by the existing network connections and the requested network connection based on the total arrival rate, the total variance and the total instantaneous variance; and

accepting the request for a network connection if a link capacity of a network node is at least as great as the determined link capacity required by the existing network connections and the requested network connection based on the total arrival rate, the total variance and the total instantaneous variance.

8. A method according to claim 7, wherein said step of summing an instantaneous variance of the change of buffer occupancy comprises the step of summing a squared coefficient of variation of the incoming traffic of each existing connection and the squared coefficient of variation of the requested network connection to provide a total squared coefficient of variation (c^2); and

wherein said step of determining the link capacity required by the existing network connections and the requested network connection based on the total arrival rate, the total variance and the total instantaneous variance comprises the step of determining the link capacity required by the existing network connections and the requested network connection based on the total arrival rate, the total variance and the total squared coefficient of variation.

9. A method according to claim 7, wherein the network is characterized by a maximum bandwidth (B) and a cell loss ratio (L) and wherein said determining step comprises the step of determining capacity (C^*) by evaluating:

$$C^*=\lambda-\beta+\sqrt{\beta^2-2\sigma^2\omega_1}$$

where

$$\beta=\frac{2B}{\alpha}\sigma^2$$

and where $\omega_1=\ln(L\sqrt{2\pi})$.

10. A method according to claim 9, wherein said step of summing an instantaneous variance of the change of buffer occupancy comprises the step of summing a squared coefficient of variation of the incoming traffic of each existing connection and the squared coefficient of variation of the requested network connection to provide a total squared coefficient of variation (c^2); and

wherein $\alpha=\lambda c^2$.

11. A method according to claim 7, wherein the network is characterized by a maximum bandwidth (B) and a cell loss

ratio (L) and wherein said determining step comprises the step of determining capacity (C) by evaluating:

$$C^*=\lambda-\beta+\sqrt{\beta^2-2\sigma^2\omega_2}$$

where

$$\beta=\frac{2B}{\alpha}\sigma^2$$

and where $\omega_2=\ln(L\lambda\sqrt{2\pi})-\ln(\sigma)$.

12. A method according to claim **11**, wherein said step of summing an instantaneous variance of the change of buffer occupancy comprises the step of summing a squared coefficient of variation of the incoming traffic of each existing connection and the squared coefficient of variation of the requested network connection to provide a total squared coefficient of variation (c^2); and

wherein $\alpha=\lambda c^2$.

13. A method according to claim **4**, wherein the network comprises an ATM network having a plurality of ATM nodes.

14. A method according to claim **13**, wherein said determining step is carried out for each node in the ATM network in a path through the network associated with the requested network connection.

15. A system for evaluating requests for a network connection, comprising:

means for determining if an acceptable cell loss ratio results from a requested network connection being added to network traffic utilizing a diffusion based statistical analysis of the network traffic; and

means for accepting the request for a network connection if an acceptable cell loss ratio results from a requested network connection being added to the network traffic based on the diffusion based statistical analysis.

16. A system according to claim **15**, wherein said means for determining comprises means for generating a diffusion based representation of the network traffic for a predefined maximum acceptable cell loss ratio resulting from the existing network connections and the requested network connection; and

wherein said means for accepting comprises means for accepting the request for network access if the diffusion based representation of the network traffic for the existing network connections and the requested network connection is less than the capacity of the network.

17. A system according to claim **16**, wherein said means for generating generates a diffusion based representation of the capacity required for existing connections and the requested connections.

18. A system for evaluating requests for a network connection, comprising:

means for determining if sufficient network capacity is available to accept the request for a network connection based upon diffusion based representations of network capacity for the existing network connections and the requested network connection; and

means for accepting the request for a network connection if sufficient network capacity is available to accept the request for a network connection based upon the diffusion based representation of the existing network connections and the requested network connection.

19. A system according to claim **18**, wherein said means for determining determines the capacity required for existing connections and the requested connections based upon a cell loss ratio (L).

20. A system according to claim **18**, wherein means for determining comprises means for generating a diffusion based representation of the network capacity required by the existing network connections and the requested network connection; and

wherein said means for accepting comprises means for accepting the request for network access if the diffusion based representation of the network capacity required by the existing network connections and the requested network connection is less than the capacity of the network.

21. A system for determining whether to accept a request for a network connection, comprising:

means for summing an arrival rate of each existing connection and the arrival rate of a requested network connection corresponding to the request for a network connection to provide a total arrival rate (λ);

means for summing a variance of arrival rate of each existing connection and the variance of arrival rate of the requested network connection to provide a total variance (σ^2);

means for summing an instantaneous variance of the change of the buffer occupancy for each existing connection so as to provide a total instantaneous variance (α);

means for determining the link capacity required by the existing network connections and the requested network connection based on the total arrival rate, the total variance and the total instantaneous variance; and

means for accepting the request for a network connection if a link capacity of a network node is at least as great as the determined link capacity required by the existing network connections and the requested network connection based on the total arrival rate, the total variance and the total instantaneous variance.

22. A system according to claim **21**, wherein said means for summing an instantaneous variance of the change of buffer occupancy comprises means for summing a squared coefficient of variation of the incoming traffic of each existing connection and the squared coefficient of variation of the requested network connection to provide a total squared coefficient of variation (c^2); and

wherein said means for determining the link capacity required by the existing network connections and the requested network connection based on the total arrival rate, the total variance and the total instantaneous variance comprises means for determining the link capacity required by the existing network connections and the requested network connection based on the total arrival rate, the total variance and the total squared coefficient of variation.

23. A system according to claim **21**, wherein the network is characterized by a maximum bandwidth (B) and a cell loss ratio (L) and wherein said means for determining comprises means for determining capacity (C*) by evaluating:

$$C^*=\lambda-\beta+\sqrt{\beta^2-2\sigma^2\omega_1}$$

where

$$\beta=\frac{2B}{\alpha}\sigma^2$$

and where $\omega_1=\ln(L\sqrt{2\pi})$.

24. A system according to claim **23**, wherein said means for summing an instantaneous variance of the change of

buffer occupancy comprises means for summing a squared coefficient of variation of the incoming traffic of each existing connection and the squared coefficient of variation of the requested network connection to provide a total squared coefficient of variation (c^2); and

wherein $\alpha=\lambda c^2$.

25. A system according to claim **21**, wherein the network is characterized by a maximum bandwidth (B) and a cell loss ratio (L) and wherein said means for determining comprises means for determining capacity (C*) by evaluating:

$$C^*=\lambda-\beta+\sqrt{\beta^2-2\sigma^2\omega_1}$$

where

$$\beta=\frac{2B}{\alpha}\sigma^2$$

and where $\omega_2=\ln(L\lambda\sqrt{2\pi})-\ln(\sigma)$.

26. A system according to claim **25**, wherein said means for summing an instantaneous variance of the change of buffer occupancy comprises means for summing a squared coefficient of variation of the incoming traffic of each existing connection and the squared coefficient of variation of the requested network connection to provide a total squared coefficient of variation (c^2); and

wherein $\alpha=\lambda c^2$.

27. A system according to claim **18**, wherein the network comprises an ATM network having a plurality of ATM nodes.

28. A system according to claim **27**, wherein said means for determining makes such determination for each node in the ATM network in a path through the network associated with the requested network connection.

29. A computer program product for evaluating requests for a network connection, comprising:

a computer-readable storage medium having computer-readable program code means embodied in said medium, said computer-readable program code means comprising:

computer-readable program code means for determining if an acceptable cell loss ratio results from a requested network connection being added to network traffic utilizing a diffusion based statistical analysis of the network traffic; and

computer-readable program code means for accepting the request for a network connection if an acceptable cell loss ration results from a requested network connection being added to the network traffic based on the diffusion based statistical analysis.

30. A computer program product according to claim **29**, wherein said computer-readable program code means for determining comprises computer-readable program code means for generating a diffusion based representation of the network traffic for a predefined maximum acceptable cell loss ratio resulting from the existing network connections and the requested network connection; and

wherein said computer-readable program code means for accepting comprises computer-readable program code means for accepting the request for network access if the diffusion based representation of the network traffic for the existing network connections and the requested network connection is less than the capacity of the network.

31. A computer program product according to claim **30**, wherein said computer-readable program code means for generating generates a diffusion based representation of the capacity required for existing connections and the requested connections.

32. A computer-program product for evaluating requests for a network connection, comprising:

a computer-readable storage medium having computer-readable program code means embodied in said medium, said computer-readable program code means comprising:

computer-readable program code means for determining if sufficient network capacity is available to accept the request for a network connection based upon diffusion based representations of network capacity for the existing network connections and the requested network connection; and

computer-readable program code means for accepting the request for a network connection if sufficient network capacity is available to accept the request for a network connection based upon the diffusion based representation of the existing network connections and the requested network connection.

33. A computer program product according to claim **32**, wherein said computer-readable program code means for determining determines the capacity required for existing connections and the requested connections based upon a cell loss ratio (L).

34. A computer program product according to claim **32**, wherein computer-readable program code means for determining comprises computer-readable program code means for generating a diffusion based representation of the network capacity required by the existing network connections and the requested network connection; and

wherein said computer-readable program code means for accepting comprises computer-readable program code means for accepting the request for network access if the diffusion based representation of the network capacity required by the existing network connections and the requested network connection is less than the capacity of the network.

35. A computer program product for determining whether to accept a request for a network connection, comprising:

a computer-readable storage medium having computer-readable program code means embodied therein, the computer-readable program code means comprising:

computer-readable program code means for summing an arrival rate of each existing connection and the arrival rate a requested network connection corresponding to the request for a network connection to provide a total arrival rate (λ);

computer-readable program code means for summing a variance of arrival rate of each existing connection and the variance of arrival rate of the requested network connection to provide a total variance (σ^2);

computer-readable program code means for summing an instantaneous variance of the change of the buffer occupancy for each existing connection so as to provide a total instantaneous variance (α);

computer-readable program code means for determining the link capacity required by the existing network connections and the requested network connection based on the total arrival rate, the total variance and the total instantaneous variance; and

computer-readable program code means for accepting the request for a network connection if a link capacity of a network link is at least as great as the determined link capacity required by the existing network connections and the requested network con-

nection based on the total arrival rate, the total variance and the total instantaneous variance.

36. A computer program product according to claim **35**, wherein said means for summing an instantaneous variance of the change of buffer occupancy comprises means for summing a squared coefficient of variation of the incoming traffic of each existing connection and the squared coefficient of variation of the requested network connection to provide a total squared coefficient of variation (c^2); and

wherein said computer-readable program code means for determining the link capacity required by the existing network connections and the requested network connection based on the total arrival rate, the total variance and the total instantaneous variance comprises computer-readable program code means for determining the link capacity required by the existing network connections and the requested network connection based on the total arrival rate, the total variance and the total squared coefficient of variation.

37. A computer program product according to claim **35**, wherein the network is characterized by a maximum bandwidth (B) and cell loss ratio (L) and wherein said computer-readable program code means for determining comprises computer-readable program code means for determining capacity (C*) by evaluating:

$$C^*=\lambda-\beta+\sqrt{\beta^2-2\sigma^2\omega_1}$$

where

$$\beta=\frac{2B}{\alpha}\sigma^2$$

and where $\omega_1=\ln(L\sqrt{2\pi})$.

38. A computer program product according to claim **37**, wherein said computer readable program code means for summing an instantaneous variance of the change of buffer occupancy comprises computer readable program code means for summing a squared coefficient of variation of the incoming traffic of each existing connection and the squared coefficient of variation of the requested network connection to provide a total squared coefficient of variation (c^2); and

wherein $\alpha=\lambda c^2$.

39. A computer program product according to claim **35**, wherein the network is characterized by a maximum bandwidth (B) and a cell loss ratio (L) and wherein said computer-readable program code means for determining comprises computer-readable program code means for determining capacity (C*) by evaluating:

$$C^*=\lambda-\beta+\sqrt{\beta^2-\sigma^2\omega_2}$$

where

$$\beta=\frac{2B}{\alpha}\sigma^2$$

and where $\omega_2=\ln(L\lambda\sqrt{2\pi})-\ln(\sigma)$.

40. A computer program product according to claim **39**, wherein said computer readable program code means for summing an instantaneous variance of the change of buffer occupancy comprises computer readable program code means for summing a squared coefficient of variation of the incoming traffic of each existing connection and the squared coefficient of variation of the requested network connection to provide a total squared coefficient of variation (c^2); and

wherein $\alpha=\lambda c^2$.

41. A computer program product according to claim **32**, wherein the network comprises an ATM network having a plurality of ATM nodes.

42. A computer program product according to claim **41**, wherein said computer-readable program code means for determining makes such determination for each node in the ATM network in a path through the network associated with the requested network connection.

* * * * *